DISSECTING THE REGULATORY ROLES AND CELLULAR FUNCTIONS OF MAMMALIAN ZSCAN5B AND PRIMATE-SPECIFIC PARALOGS

BY

YOUNGUK SUN

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Cell and Developmental Biology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2016

Urbana, Illinois

Doctoral Committee:

Professor Lisa J. Stubbs, Chair
Associate Professor David H. Rivier
Associate Professor Craig A. Mizzen
Assistant Professor Eric C. Bolton
Assistant Professor Rachel Smith-Bolton

ABSTRACT

Certain subfamilies of zinc finger transcription factors (ZNF-TFs), especially those including KRAB or SCAN protein-interaction domains, have evolved rapidly in mammals due to repeated rounds of gene duplication, deletion, and divergence. The functions of the encoded proteins, and the evolutionary role of their duplication and divergence, are not well understood. I have focused on mammalian ZNF subfamilies that are rooted by relatively unique, conserved members but that have duplicated to generate groups of divergent paralogous transcription factors in primate species. One such family is the ZSCAN5 subfamily, represented by a single, conserved founding member, ZSCAN5B, and clustered paralogs that arose early in primate history, and this clustered subfamily has been the subject of my thesis work. I have shown that three of the primate paralogs, ZSCAN5A, B, and D, are expressed in different tissues and cell types while ZSCAN5C is not detectably expressed in most tissues. Using chromatin immunoprecipitation (ChIP-seq) in human cell lines and mouse placenta tissue, I have identified mammalian ZSCAN5B as a protein that binds with high specificity to tRNA genes (tDNA) and other polymerase III (Pol III) transcripts including several types of transposable elements (TEs). Primate-specific paralogs ZSCAN5A and ZSCAN5D also bind tDNAs, although ZSCAN5D preferentially recognizes a subset of TE-derived "extra-TFIIIC" or ETC sites.

I also used siRNA knockdown followed by transcriptome sequencing (RNA-seq) to identify biological functions of the dominantly expressed paralog *ZSCAN5A* and the ancestral *ZSCAN5B* gene. In addition to tDNAs and other Pol III transcripts, I show that *ZSCAN5A* and *ZSCAN5B* gene knockdown also dysregulates many Pol II genes including some that are located near the DNA binding regions. Significant overlap between the differentially expressed gene (DEG) sets detected after *ZSCAN5A* or *ZSCAN5B* knockdown suggested that the two proteins may act cooperatively through their similar SCAN domains. The DEG sets thus suggested similar biological functions for ZSCAN5A and ZSCAN5B, but also highlighted some potentially distinct activities. In particular, based on these data I hypothesize that ZSCAN5B primarily regulates ribosome biogenesis and tRNA processing, whereas ZSCAN5A has an especially important role in

mitotic progression. Consistent with this latter function, I show that ZSCAN5A as well as its human paralogs are cell-cycle regulated with peak expression around the time of mitosis.

To test these hypotheses about gene function, I generated a series of resources including both human cell lines and transgenic mice engineered to overexpress or knock down *ZSCAN5* genes specifically. Focusing on the predicted role of ZSCAN5A in cell cycle progression, I showed that *ZSCAN5A* knockdown dysregulates cell cycle progression in cultured human cells, leading to an accumulation of cells in mitotic phase and the appearance of aneuploid cells. These data suggest that, either through a cyto-architectural function, regulation of Pol III transcription, position effects on Pol II genes, or a combination of these roles, ZSCAN5A has an essential role in maintenance of chromosome integrity in human cells. These data open up many new avenues for future exploration.

Together these data define unexpected functions for a conserved mammalian gene and two paralogs found only in primate genomes. The human paralogs have evolved from the basic tDNA-binding function of the conserved founder gene, *ZSCAN5B,* to bind more widely to Pol III transcripts and other TFIIIC binding sites and affect the transcription of a wider selection of nearby genes. Data presented here indicate that despite being duplicated only in primates, *ZSCAN5A* has been integrated into the control of a very ancient process, namely chromosome segregation during mitosis, and plays an important role in chromosome integrity. The new tools and resources I have developed provide essential tools needed to explore those functions in further depth.

ACKNOWLEDGEMENTS

Words can only say so much how much I appreciate all the support and encouragement from my advisor, Dr. Lisa Stubbs, my colleagues and friends, and my family. Firstly, this project would have not happened and done without Dr. Lisa Stubbs and thesis committee members' enthusiastic support and untiring guidance throughout the completion of my degree. I am also very lucky to have my friend, Dr. Chase Bolt, and grateful for the countless scientific discussions and moral support over the countless pints of beers. Dr. Elbert Branscomb's inspirational input, Christopher Seward and Joe Troy's proficient bioinformatics analysis, Huimin Zhang and Jaeeun Yoo's remarkable experimental assistances have greatly fueled the completion of this study. I am much indebted to my parents, Yongseung Sun and Kyungran Baik, and my sister, Youngeun Sun, for their incessant and motivating encouragement to keep me moving forward. Many thanks are due to my closest and loving friends – Jiyoung Jeong, Dongwook Kim, Jihye Nam – and colleagues, Dr. Patricia Annie Weisner, Li-Hsin Chang, Erin Borchardt, and Seeon Lee. Lastly, I would like to express my sincere gratitude to my undergraduate advisor, Dr. Suk-Tae Kwon, for having faith in me and gave me an opportunity to start my career in the field of molecular biology.

TABLE OF CONTENTS

CHAPTER 1:

INTRODUCTION

Evolution of Regulatory Networks

While it is still controversial and vague to define a true meaning of 'life', scientifically and traditionally, 'life' is specifically distinguished by the capacity that it can grow, metabolize, respond to stimuli, adapt, and reproduce. This reproductive capacity elicits numerous generations of offspring, and each generation offers a new opportunity for evolution of form and function. From simple single-cell bacteria to highly complex vertebrate mammals, every known living organism on the planet earth follows a simple rule known as the "Central dogma of molecular biology"(Crick, 1970; Crick, 1958). Each organism carries its own genetic material that is transcribed and further translated into useful peptides to sustain its viability, and replicated to produce additional progeny. Amongst these important processes, an accurate and sophisticated transcription of each gene is essential for the successful development and survival of the living organism (Jacob and Monod, 1961). This fine control is called "transcriptional regulation".

The transcriptional regulation system is principally composed of *cis*- and *trans*-regulatory elements (Davidson 2006). *Cis*-regulatory elements are typically non-coding DNA segments that contain binding sites for transcription factors (TFs) and other regulatory proteins that are required to activate and sustain transcription; the well-characterized examples of these *cis*-elements are enhancers and promoters (Blackwood and Kadonaga, 1998; Pennacchio et al., 2013; Smale and Kadonaga, 2003). Conversely, *trans*-regulatory factors include proteins, RNAs, and other diffusible molecules that affect gene expression, primarily by interacting with the *cis* DNA sites (Gilad et al., 2008). The basic biochemical function of a TF is firstly, to recognize and bind a short, specific string of nucleotides (called a motif) within *cis*-regulatory regions, and secondly, to recruit or bind other proteins relevant to transcriptional regulation; these include other

TFs, chromatin remodeling complex components, and the general RNA polymerase machinery (Karin, 1990; Latchman, 1997). Thus, the interaction between a TF and a specific *cis*-regulatory element may endow positive or negative changes to the basal expression of a target gene (Lee and Young, 2000; Nikolov and Burley, 1997; Roeder, 1996).

Because these systems orchestrate the specification of a particular cell type or structure, heritable changes to regulatory networks result in the evolution of animal morphology and phenotypic variations across organisms (Carroll et al., 2013; Rubinstein and de Souza, 2013). Throughout the course of evolution, as biological systems have become more complex, the amount of information devoted to the transcription regulation must increase accordingly. Or conversely, it is thought that the expansion and variations in regulatory systems may have facilitated the emergence of complex organisms and the evolution of multicellularity (Levine and Tjian, 2003).

There still is an ongoing debate whether the changes to *cis*-regulatory elements or to *trans*-regulatory elements have served as the prevalent means by which regulatory networks evolve. However, available evidence suggests that evolutionary changes in both *cis*- and *trans*- components have played important roles. Comparative genomics studies have suggested a substantial level of *cis*-regulatory variation across species, and it has long been hypothesized that these changes in *cis*-regulatory elements are predominantly responsible for the changes in regulatory network function (Borneman et al., 2007; Bradley et al., 2010; Carroll, 2005; Odom et al., 2007; Prud'homme et al., 2007; Wray, 2007). Other support for this arises because of the likely negative pleiotropic effects that could occur when mutations in protein-coding regions of TFs take place (Carroll, 2005; Hsia and McGinnis, 2003). In other words, variation in a transcriptional regulator can affect multiple target genes simultaneously and result in widespread detrimental ramifications, whereas, a mutation in single a *cis*-regulatory element will only affect changes in the local gene expression pattern (Carroll, 2005; Davidson and Erwin, 2006; Doebley and Lukens, 1998; Hsia and McGinnis, 2003; Wittkopp, 2005; Wray, 2007).

However, recent studies have accumulated significant evidence that variations in *trans*-regulatory proteins may be more common than it was originally appreciated (Brem et al., 2002; Bustamante et al., 2005; Yvert et al., 2003). Altering the expression, post-

translational modification, and molecular interactions of regulatory proteins have been suggested as strategies for promoting regulatory network evolution (Grove et al., 2009; Lynch et al., 2011; Reece-Hoyes et al., 2013; Sayou et al., 2014).

Expansion of TFs and C2H2 zinc finger proteins during evolution

At the base of the metazoan lineage, many TF families have undergone a series of duplications and divergence resulting in numerous homologs, which provide a valuable source of novel material for building extended gene regulatory networks (Degnan et al., 2009; Pérez et al., 2014; Teichmann and Babu, 2004).

There are two primary ways in which new TF homologs can be generated. When species diverge from a common ancestor, each initially retains nearly identical sets of TFs, or one-to-one orthologs. Until a novel regulatory mechanism is invented, the orthologous proteins tend to serve the same function in each species as they did in the ancestor. These newly generated orthologs exist under a great amount of evolutionary constraint to remain conserved. On the other hand, TFs generated by gene duplication events, which create new paralogs, have ample freedom to evolve (Hoekstra and Coyne, 2007). While most gene duplicates are simply lost, others undertake some of the ancient roles of the original TF by reducing the load on each copy and giving each copy more plasticity to change. This subdividing is known as “subfunctionalization” (Force et al., 1999). Ultimately, if one copy retains all of the ancestral roles, the other paralogs will have essentially no constraint and can acquire a new function, or neofunctionalize (Reece-Hoyes et al., 2013). In this way, generation of parallels results in modularity within a TF family, because each parallel endows the others with greater freedom to change.

Of the various families of TFs, the C2H2 zinc finger (ZNF) TF gene family is by far the largest, and the most evolutionarily divergent (Huntley et al., 2006; Lander et al., 2001; Urrutia, 2003; Venter et al., 2001). ZNF genes encode one of the most abundant protein families in all eukaryotic species, being found, for example, in approximately 2% of all human genes (Lander et al., 2001). Several subfamilies of this large ZNF family have evolved through massive lineage-specific expansions, thus providing a putative

source for gene regulatory change (Huntley et al., 2006; Liu et al., 2014; Nowick et al., 2011; Nowick and Stubbs, 2010; Stubbs et al., 2011) (*Fig. 1.1*).

The modular structure of the ZNF genes and proteins may be especially suited to modifications that lead to evolutionary change. ZNF TFs contain at least 3, and as many as 40 discrete C2H2 zinc finger motifs that function cooperatively, although each contributes distinctly and additively to DNA binding properties of the protein (Huntley et al., 2006; Iuchi, 2001). Within each motif, conserved cysteine and histidine amino acids coordinate binding to a single zinc ion; zinc binding allows the finger to fold into a compact structure containing a beta-turn that includes the conserved cysteines and an alpha-helix that includes the conserved histidines (Frankel et al., 1987; Klug and Schwabe, 1995; Pavletich and Pabo, 1991; Stubbs et al., 2011) (*Fig. 1.2A*). Even though there is variation between some amino acid positions that compose individual fingers, C2H2 motifs are distinguished by the highly conserved, repetitive amino acid pattern [TGEKP(Y/F)] that links each finger together and other relatively conserved residues within the structure (Bellefroid et al., 1989). Each individual finger recognizes 3 base pairs of DNA through interactions between the α-helix region and the DNA site, in such a way that the array of tandem fingers winds around the major groove of the DNA molecule in a highly sequence-specific manner (Pavletich and Pabo, 1991) (*Fig. 1.2B*). Within each finger, four specific DNA-contacting residues, at amino acid positions -1, +2, +3 and +6 of the alpha helix, are primarily responsible for contacting DNA (Klug, 2010; Wolfe et al., 2000); we have referred to the pattern of four DNA-contacting amino acids in each zinc finger as a protein's "fingerprint". These fingerprint patterns can be compared across species to determine the conservation of a ZNF-DNA binding (Stubbs et al., 2011).

History and known functions of SCAN-domain ZNF genes

Usually found on the N-terminus of the C2H2 zinc finger elements in some of the transcription factors, a series of extended sequence motifs are associated to help define these regulators. These structural modules are known to regulate subcellular localization, DNA binding, and gene expression by controlling selective recruitment of the TFs or

other chromosome remodeling proteins. These associated modules include the *Krüppel*-associated box (KRAB) (Bellefroid et al., 1991), the poxvirus and zinc finger (POZ) domain (Bardwell and Treisman, 1994), which is also known as the BTB domain (Broad-Complex, Tramtrack, and Bric-a-brac) (Albagli et al., 1995), and the SCAN domain (Williams et al., 1995). These domains define subgroups within the C2H2 family and may provide insights into the functions of the members of this large family of zinc finger TFs.

The SCAN domain, which is also known as the leucine rich region, is a highly conserved 84-residue motif that is found near the N-terminus of a subfamily of C2H2 zinc finger proteins (Pengue et al., 1994). The SCAN domain was originally identified in ZNF174 (Williams et al., 1995). The name was based on the first letters found in some of the founding members of the family (SRE-ZBP, Ctfin51, AW-1 (ZNF174), and Number 18). Studies utilizing either mammalian or yeast two-hybrid systems demonstrated the protein interaction function of the SCAN domain, mediating self-association or selective association with other proteins (Sander et al., 2000; Schumacher et al., 2000; Williams et al., 1995). However, unlike the KRAB and BTB/POZ domains, no apparent transcriptional activation or repression properties were detected for SCAN domains (Sander et al., 2000; Williams et al., 1999). Affinity purified SCAN domains of ZNF174 exhibited stable homodimerization with high melting temperature *in vitro* (Stone et al., 2002). In view of the regulatory repertoire, TF dimerization can expand the selectivity of protein-DNA interactions, by increasing the length of the DNA motif that interacts with the proteins as they bind to DNA side-by-side; furthermore, since most TFs can dimerize with several others in different combinations, this process can also generate increased diversity of genomic binding sites from a relatively small number of proteins (Lamb and McKnight, 1991). Indeed, the SCAN domains found in lower vertebrates are not associated with C2H2 DNA-binding motifs, and it is thought that the rapid and lineage-specific expansion of SCAN-ZNF proteins once this combination arose may have contributed to the diversity seen in higher vertebrates (Sander et al., 2003).

At present, only a few functional descriptions of SCAN-ZNFs have been reported. Zscan10 (Zfp206), a SCAN-ZNF with 14 tandem C2H2 ZNFs, is known to maintain embryonic stem cell (ESC) pluripotency by interacting with the established pluripotency

marker Sox2 and Oct4 (Wang et al., 2007). Kraus and colleagues have suggested a role for Zscan10 in maintaining the progenitor cell subpopulation and/or impacting fate choice decisions, which are based on the findings that Zscan10 homozygous null mice exhibited reduced weight, mild hypoplasia in the spleen, heart and long bones and phenocopy an eye malformation described for Sox2 hypomorphs (Kraus et al., 2014). Another SCAN-ZNF protein, Zscan4, has also been linked to early steps in stem cell differentiation. *Zscan4* gene expression was shown to reactivate early embryonic genes, enhance the efficiency of generating induced pluripotent stem cell (iPSC), and functionally replace Myc, which increases cell proliferation and suppresses genome stability (Hirata et al., 2012; Nakagawa et al., 2010).

Long-range gene regulation and RNA polymerase III transcription system

Conventional concepts of gene regulation involve one or more TFs that bind to *cis*-regulatory elements, usually a promoter, to either activate or repress the nearby gene (Lee and Young, 2000; Nikolov and Burley, 1997; Roeder, 1996). However, studies over the last twenty years have suggested additional models, in which gene regulation can be controlled by regulatory elements located some distance from the transcription start site (TSS) (Blackwood and Kadonaga, 1998; Pennacchio et al., 2013; Smale and Kadonaga, 2003). These elements have been proposed to interact with gene promoters through chromatin loops, organized by proteins that mediate these higher order chromatin configurations and nuclear architectures by binding to specific DNA sites (Clelland and Schultz, 2010; Conesa et al., 2005; Simms et al., 2008). There are very likely many different mechanisms that form and mediate these structures, and the classes of proteins involved in these processes are just beginning to be understood. However, one system of *cis*- and *trans*- elements that mediate chromatin architecture is particularly ancient, being found in all eukaryotic genomes examined to date (Pascali and Teichmann, 2013). This system involves RNA polymerase III (Pol III) transcription units and the Pol III TF, TFIIIC.

Among the three eukaryotic DNA-dependent transcription complexes, RNA Pol III and was discovered as the complex responsible for production of transfer RNA (tRNA)

and a limited number of other small RNAs including the small 5S ribosomal RNA (rRNA), the 7SL RNA, U6 spliceosomal RNA (Geiduschek and Kassavetis, 2001; Huang and Maraia, 2001; Paule and White, 2000). Mammalian RNA Pol III genes are transcribed from three distinct types of promoters; type 1 (5S genes) and type 2 (tRNAs and viral associated RNA) promoters are internal to the genes, while type 3 (U6, RNase P, RNase MRP genes) promoters are located upstream of TSS (Dieci et al., 2007; Orioli et al., 2012). Type 1 and 2 promoters include the A and the B box motifs, which recruit the transcription factor TFIIIC. The B box has been shown to primarily recruit TFIIIC and B Box sequences with robust TFIIIC binding also occur at sites that are not actively transcribed by Pol III. These sites, known as 'Extra-TFIIIC' (ETCs) sites exist in every species that have been examined and have been referred to as 'Chromosome Organizing Clamps' (COCs) in yeast (Moqtaderi and Struhl, 2004; Noma et al., 2006).

Some of these ETC sites are very likely created by transposable elements. Indeed many of the most common repetitive elements in higher eukaryote genomes are derived from Pol III genes. For example, small interspersed nuclear elements (SINEs), which represent up to 11% of the entire human genome (Cordaux and Batzer, 2009), have been suggested to originate from 7SL RNA genes, and a more ancient element, called MIR, is thought to originate from tRNA genes (tDNAs) (Jurka et al., 1995; Silva et al., 2003; Smit and Riggs, 1995). Many of these SINE elements retain a recognizable B Box motif and TFIIIC binding, and a few of these elements are actively transcribed by Pol III (Aleman et al., 2000; Berger and Strub, 2011; Cordaux and Batzer, 2009; Shaikh et al., 1997).

Pol III transcription is known to be critical for cell survival, but in addition to its canonical tRNA synthesis activity, recent studies demonstrate other crucial roles in various nuclear processes, including effects on nucleosome positioning, global genome and subnuclear organization, and direct effects on RNA pol II transcription (Clelland and Schultz, 2010; Conesa et al., 2005; Simms et al., 2008). In fact, tDNAs have been shown to act as insulators in both yeast and human cells (Ebersole et al., 2011; Hull et al., 1994; Kendall et al., 2000; Kinsey and Sandmeyer, 1991; Raab et al., 2012). For example, tRNA gene-mediated silencing (tgm) and other effects on nearby gene expression may involve an enhancer-blocking insulator activity facilitated by chromosome loop

formation. In yeast, tgm silencing was first discovered when it was shown that active transcription of tRNA genes represses Pol II transcription from nearby promoters (Bolton and Boeke, 2003; Hull et al., 1994; Kendall et al., 2000; Kinsey and Sandmeyer, 1991). Also in *S. cerevisiae*, widely distributed tRNA genes have been shown to cluster in the proximity of 5S RNA genes in the nucleolus (Thompson et al., 2003; Wang et al., 2005) and this unique tDNA clustering was found crucial for the inhibition of Pol II transcription (Haeusler et al., 2008; Wang et al., 2005). The nucleolar relocation of tDNA clusters results in the formation of chromatin loops and this is facilitated by cooperative interactions between TFIIIC and condensin (D'Ambrosio et al., 2008; Haeusler et al., 2008; Wang et al., 2005).

However, the first type of insulator function assigned to tDNA was heterochromatin barrier activity. Specifically, a $tRNA^{Thr}$ gene located near the centromere was shown to restrict the spread of heterochromatin in *S. cerevisiae* (Donze et al., 1999; Donze and Kamakaka, 2001). Even though, mechanisms for establishing barrier activity vary in different tDNAs, it has been proposed that TFIIIB, which is recruited by and after TFIIIC, attracts histone acetyltransferase and other chromatin remodeling complexes, such as ISW2, to demarcate the region (Bachman et al., 2005; Donze and Kamakaka, 2001; Gelbart et al., 2005). While most work has focused on the yeasts including *S. cerevisiae* and *S. pombe*, several lines of evidence suggested that mammalian Pol III complexes exert similar functions, but with a few potentially important differences (Van Bortle and Corces, 2012).

In mammalian cells, the protein encoded by the proto-oncogene *c-MYC* was found to activate Pol III transcription not only by directly interacting with TFIIIB (Gomez-Roman et al., 2003), but also by recruiting chromatin modifying complexes, including histone acetyltransferase (KAT2A) (Flinn et al., 2002) and TRRAP (Kenneth et al., 2007) to mobilize histones around tDNA regions to facilitate TFIIIC binding. In turn, TFIIIC appears to introduce active histone modifications when bound at ETC sites near Pol II genes, but the opposite at the transcribed Pol III genes (Barski et al., 2010; Jin et al., 2009; Oler et al., 2010). Moqtaderi and colleagues exploited ChIP-seq in human cells to also discover that 1) A/B box regions are free of nucleosomes; 2) most ETCs were located <1 kb of Pol II promoters; 3) TFIIIC, but not Pol III, binds to ETC and co-

localizes with CCCTC binding factor (CTCF) at those sites, and 4) tDNAs and ETC sites interact with each other preferentially across long-range chromatin loops (Moqtaderi et al., 2010). These authors also showed, using a reporter assay system that tDNA sequences could function as enhancer-blocking insulators in cultured mammalian cells (Moqtaderi et al., 2010). The link between TFIIIC and CTCF binding is of particular interest, since most mammalian insulators bind CTCF and CTCF-cohesin complex is required for enhancer-blocking insulation in vertebrates (Bell and Felsenfeld, 1999; Filippova et al., 2005; Magdinier et al., 2004; Mishiro et al., 2009; Phillips and Corces, 2009; Wendt et al., 2008).

Interestingly, SINEs, which make up a substantial fraction of the mammalian ETC repertoire, have also been linked to enhancer-blocking insulation (Lunyak et al., 2007; Román et al., 2011). Alu elements are often associated with cohesin and CTCF homodimer or multimers to result in the formation of interchromosomal loops (Hakimi et al., 2002; Parelho et al., 2008; Wendt et al., 2008; Wood et al., 2010). Alu RNAs are normally repressed of their expression by nucleosome positioning and DNA methylation (Englander et al., 1993; Kim et al., 2001; Li et al., 2000; Liu et al., 1994; Liu and Schmid, 1993; Russanova et al., 1995)

In this context, it is interesting to note that there has been approximately 250-fold expansion in the size of genome from yeast to human, yet that the number of tDNAs has barely increased (274 in yeast versus approximately 500 in the human genome) (Canella et al., 2010; Dittmar et al., 2006; Harismendy et al., 2003; Moqtaderi and Struhl, 2004; Roberts et al., 2003). The numbers of tDNAs is probably constrained to some degree by their participation in other cellular functions. Pascali has suggested that the Pol III-derived repetitive elements such as SINEs may serve to compensate for the relatively low density of tDNAs in mammals, and that they may play a significant role in organizing three-dimensional structure of the genome in higher eukaryotes (Pascali and Teichmann, 2013).

Several lines of evidence now support this interesting hypothesis. If it is correct, it seems likely that the genomes of higher eukaryotes must also have developed novel mechanisms for regulating these elements, and their chromatin-organizing functions. Given their evolutionary history, one might predict that certain members of the KRAB-

or SCAN-ZNF families might be involved in these novel and species-diverse functions. Indeed, our studies suggest a role for one SCAN-ZNF subfamily, which is the primary subject of this thesis.

Zscan5b and the ZSCAN5 family

The murine Zscan5b protein contains five C2H2 zinc-finger domains (sequence-specific DNA binding motif) and SCAN domain (effector) at its N-terminus (*Fig. 1.3A*). While murine *Zscan5b* stands out as a single, highly conserved gene in mouse and other mammalian genomes, it has recently duplicated to generate a cluster of four primate-specific daughter genes, the ZSCAN5 family, *ZSCAN5A/B/C/D* (Huntley et al., 2006) (*Fig. 1.3B*). These human orthologs have been recently annotated and are shown to encode full-length SCAN-ZNF proteins. Although murine Zscan5b and human ZSCAN5B are quite divergent, sharing only 70% amino acid identity in their finger domains, these proteins share almost identical amino acid sequences in the DNA binding residues (-1, +2, +3, +6) of each finger suggesting that they may recognize similar DNA binding sites and targets (*Fig. 1.3C*). However, the sequence differences among the three other human paralogs suggest that they may bind to related but distinct DNA binding motifs. While, mouse, marmoset and human ZSCAN5B proteins share almost identical fingerprints, the three primate-specific paralogs have diverged from the ancestral gene in fingerprint patterns (*Fig. 1.3C*). After an initial phase of divergence in early primate history, these patterns have been very well conserved. In contrast to the zinc fingers, the SCAN domains of the four human ZSCAN5 family members are nearly identical (95-98% identity between members; not shown). Given that the SCAN domains mediate protein dimer formation (Sander et al., 2000; Schumacher et al., 2000; Williams et al., 1995), ZSCAN5 family members may interact as heterodimers, cooperating in different combinatorial patterns in the tissues and cell types in which they are co-expressed.

In this dissertation, I attempt to address questions such as: What kinds of functions, and which pathways, have acquired novel regulatory functions through duplicative expansion of the SCAN-ZNF transcription factor families? Do the novel proteins take on properties that are distinct from conserved parental genes, or do they “subfunctionalize”,

taking on a subset of roles that were originally carried out by the mammalian ancestors?

Chapter 2 covers the expression profile, identification, and validation of DNA-binding sites of ZSCAN5 family. The human paralogs have diverged from *ZSCAN5B* in both zinc-finger structure and expression patterns, but share expression in dividing cell populations with a distinct peak around the time of mitosis. Our data have revealed that tDNAs are the preferred binding sites for mouse and human ZSCAN5B as well as ZSCAN5A; paralog ZSCAN5D also binds similar loci but preferential binding to MIR, SINE, and LINE2 elements including a previously described novel ETC motif (Moqtaderi et al., 2010). We also examine the differential expression of nearby Pol II genes to understand the regulatory functions of ZSCAN5 family.

Chapter 3 further investigates the biological functions of ZSCAN5s with regard to their roles in the cell cycle progression using overexpression and knockdown transgenic human cell lines (*in vitro*) and transgenic mice (*in vivo*). In addition, to comprehend the mechanisms of ZSCAN5 paralogs, we have determined their cellular localization during specific cell cycle stages and examine putative combinatorial interactions between the paralogs and with other predicted binding partners.

Chapter 4 describes the generation of some essential tools, including expression constructs and two sets of transgenic mouse lines developed to address the root biological functions of the ZSCAN5 family through analysis of the unique mouse family member, *Zscan5b*.

FIGURES

Figure 1.1.

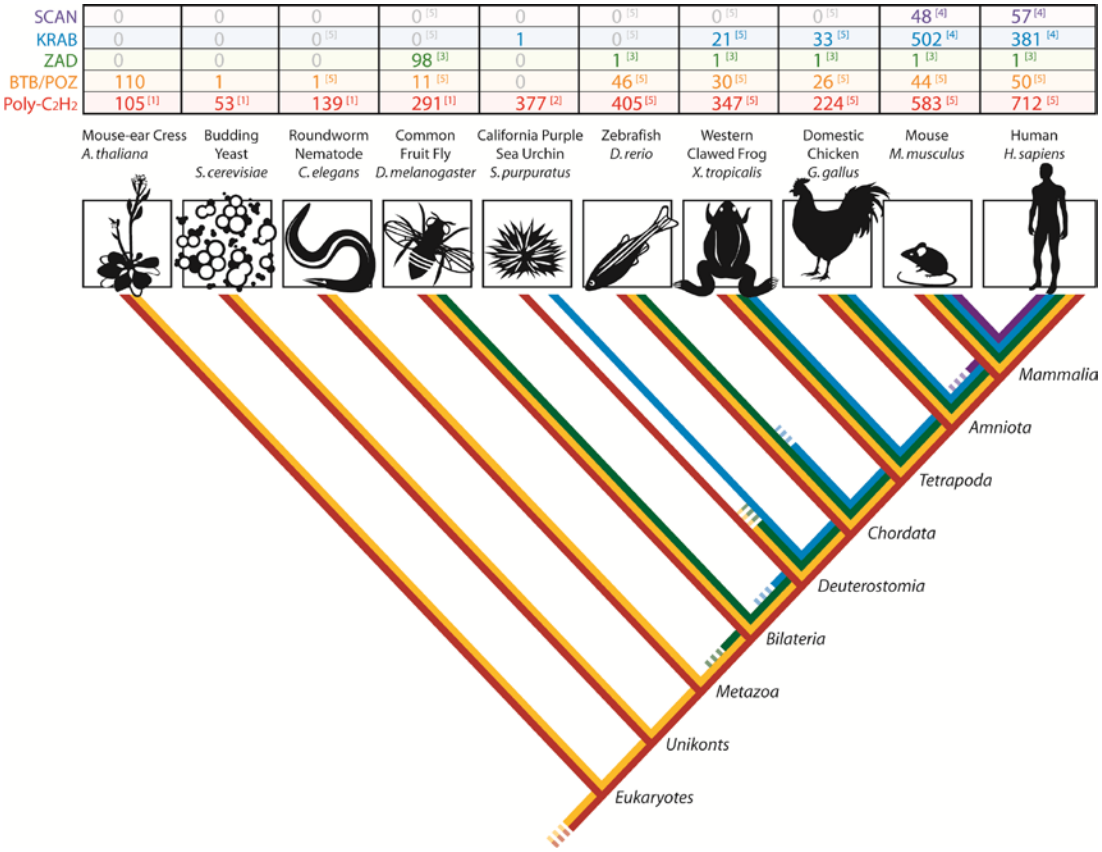


Figure 1.1. (cont.) Phylogenetic distribution of C2H2 ZNF protein families

The distribution and number of poly-C2H2 ZNF proteins in different families is shown in this phylogeny of all eukaryotic model systems. The gain and loss of genes over evolution can be seen along the tree for all poly-C2H2 (red) BTB/POZ (orange), ZAD (green), KRAB (blue), and SCAN (purple) KZNF families. The numbers of genes in each family are shown in the accompanying table. Information was compiled from the PFAM Database (Finn *et al.* 2010), unless otherwise noted as coming from [1] (Riechmann et al. 2000),[2] (Materna *et al.* 2006), [3] (Chung *et al.* 2002), [4] (Huntley et al. 2006) or [5] (Emerson & Thomas 2009). Modified from Stubbs, Sun and Caetano-Anolles, 2011

Figure 1.2.

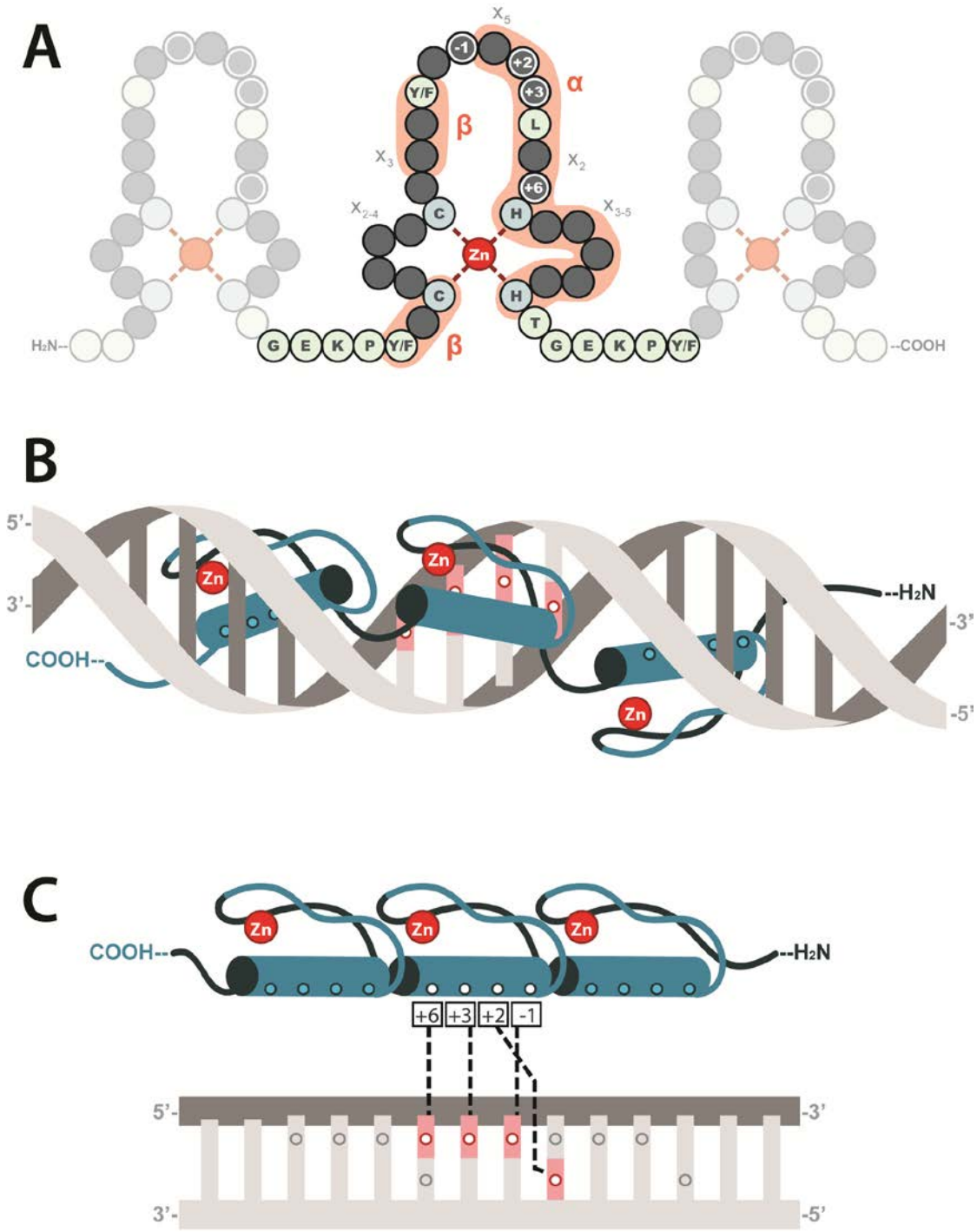

A
X5
-1
+2
+3
α
Y/F
β
L
X3
X2
+6
X2-4
C
H
X3-5
Zn
C
H
T
H2N--
G E K P Y/F
β
G E K P Y/F
--COOH
B
5'-
3'-
COOH--
Zn
Zn
Zn
--H2N
-3'
-5'
C
COOH--
Zn
Zn
Zn
--H2N
+6
+3
+2
-1
5'-
-3'
3'-
-5'

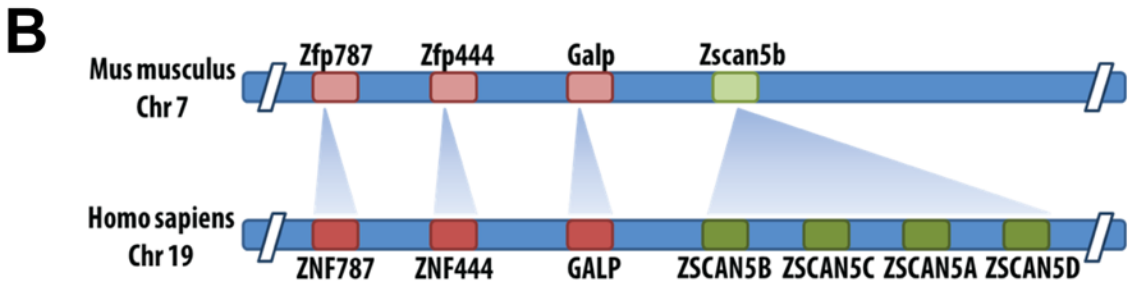
Figure 1.2. (cont.) (A) Structure of tandem C2H2 DNA-binding motif

Individual zinc ions interact with paired cysteine and histidine residues, stabilizing two β-sheets and one α-helix (as indicated by the areas within the dotted lines) into the recognizable "finger" structure shown here, with the latter containing the DNA-binding interface as indicated in the figure at the –1, +2, +3, and +6 positions relative to the helix. The structure of a finger sequence motif is represented, with *X* denoting an amino acid residue of any type with the subscript representing the number (ie. $X_{2\text{-}4}$ represents a chain of between 2 and 4 non-specified amino acid residues). The consensus sequence – *TGEKP(Y/F)* – is a highly conserved "H/C link" region between consecutive fingers.

(B), (C) DNA binding properties of tandem C2H2-ZNF domains

The α-helices of C2H2-ZNF motifs contain amino acid residues that bind to DNA nucleotides at the -1, +2, +3, and +6 sites represented as white dots on the helices of a set of tandem zinc fingers. The relationship between fingers and nucleotides is not one-to-one, as the amino acid at the *+2* position will interact with the nucleotide complementary to the neighboring finger's *+6* binding site. In this fashion, fingers wind around the major groove of the DNA molecule. (Stubbs *et al.* 2011)

Figure 1.3.



C

Protein	DNA binding amino acid residues (-1/2/3/6) of zinc finger domains				
	1st	2nd	3rd	4th	5th
Zscan5b	Y S R L	Q S D V	H S T G	H G N V	Q G T R
ZSCAN5B	Y S Q L	Q S D V	H S T G	H G N V	Q G T R
ZSCAN5B	Y S Q L	Q S D V	H S T G	H G N V	Q G T R
ZSCAN5A	C S K I	Q I S F	Q S Y C	Y G S E	R K L R
ZSCAN5A	C S K I	Q I S F	Q S Y C	Y G S E	R K L R
ZSCAN5C	Y G K I	Q I G L	Q S Y C	Y G S E	R A T R
ZSCAN5C	Y G K V	Q I G F	Q S Y C	Y A N E	R A T R
ZSCAN5D	Y S K I	Q S D V	R F S C	Y K N Q	R E T R
ZSCAN5D	C S K I	Q S D V	R F S C	Y K N E	R E T Y

Figure 1.3. (cont.)

(A) Structure of Zscan5b protein

A SCAN (protein interaction motif) domain tethered to N-terminus to five tandem C2H2 zinc finger DNA binding domains.

(B) Gene duplication of *Zscan5b*

While mouse *Zfp787*, *Zfp444*, *Galp* have 1:1 orthologs in human chromosome 19, *Zscan5b* has been duplicated to create 4 primate-specific duplicates, *ZSCAN5A/B/C/D.*

(C) ZSCAN5 family fingerprints comparison

Amino acids that are positioned at -1, 3, 6 are responsible for binding to specific DNA sequences. Mouse, marmoset and human ZSCAN5B proteins share almost identical -fingerprints indicating that they bind to very similar DNA motifs. In contrast, the three primate-specific paralogs have diverged significantly from the parental gene and from each other in fingerprint patterns after duplication.

REFERENCES CITED

1. Albagli, O., Dhordain, P., Deweindt, C., Lecocq, G., and Leprince, D. (1995). The BTB/POZ domain: a new protein-protein interaction motif common to DNA- and actin-binding proteins. Cell Growth Differ *6*, 1193-1198.

2. Aleman, C., Roy-Engel, A.M., Shaikh, T.H., and Deininger, P.L. (2000). Cis-acting influences on Alu RNA levels. Nucleic Acids Res *28*, 4755-4761.

3. Bachman, N., Gelbart, M.E., Tsukiyama, T., and Boeke, J.D. (2005). TFIIIB subunit Bdp1p is required for periodic integration of the Ty1 retrotransposon and targeting of Isw2p to S. cerevisiae tDNAs. Genes Dev *19*, 955-964.

4. Bardwell, V.J., and Treisman, R. (1994). The POZ domain: a conserved protein-protein interaction motif. Genes Dev *8*, 1664-1677.

5. Barski, A., Chepelev, I., Liko, D., Cuddapah, S., Fleming, A.B., Birch, J., Cui, K., White, R.J., and Zhao, K. (2010). Pol II and its associated epigenetic marks are present at Pol III-transcribed noncoding RNA genes. Nat Struct Mol Biol *17*, 629-634.

6. Bell, A.C., and Felsenfeld, G. (1999). Stopped at the border: boundaries and insulators. Curr Opin Genet Dev *9*, 191-198.

7. Bellefroid, E.J., Lecocq, P.J., Benhida, A., Poncelet, D.A., Belayew, A., and Martial, J.A. (1989). The human genome contains hundreds of genes coding for finger proteins of the Kruppel type. DNA *8*, 377-387.

8. Bellefroid, E.J., Poncelet, D.A., Lecocq, P.J., Revelant, O., and Martial, J.A. (1991). The evolutionarily conserved Kruppel-associated box domain defines a subfamily of eukaryotic multifingered proteins. Proc Natl Acad Sci U S A *88*, 3608-3612.

9. Berger, A., and Strub, K. (2011). Multiple Roles of Alu-Related Noncoding RNAs. Prog Mol Subcell Biol *51*, 119-146.

10. Blackwood, E.M., and Kadonaga, J.T. (1998). Going the distance: a current view of enhancer action. Science *281*, 60-63.

11. Bolton, E.C., and Boeke, J.D. (2003). Transcriptional interactions between yeast tRNA genes, flanking genes and Ty elements: a genomic point of view. Genome Res *13*, 254-263.

12. Borneman, A.R., Gianoulis, T.A., Zhang, Z.D., Yu, H., Rozowsky, J., Seringhaus, M.R., Wang, L.Y., Gerstein, M., and Snyder, M. (2007). Divergence of transcription factor binding sites across related yeast species. Science *317*, 815-819.

13. Bradley, R.K., Li, X.Y., Trapnell, C., Davidson, S., Pachter, L., Chu, H.C., Tonkin, L.A., Biggin, M.D., and Eisen, M.B. (2010). Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species. PLoS Biol *8*, e1000343.

14. Brem, R.B., Yvert, G., Clinton, R., and Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. Science *296*, 752-755.

15. Bustamante, C.D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M.T., Glanowski, S., Tanenbaum, D.M., White, T.J., Sninsky, J.J., Hernandez, R.D., *et al.* (2005). Natural selection on protein-coding genes in the human genome. Nature *437*, 1153-1157.

16. Canella, D., Praz, V., Reina, J.H., Cousin, P., and Hernandez, N. (2010). Defining the RNA polymerase III transcriptome: Genome-wide localization of the RNA polymerase III transcription machinery in human cells. Genome Res *20*, 710-721.

17. Carroll, S.B. (2005). Evolution at two levels: on genes and form. PLoS Biol *3*, e245.

18. Carroll, S.B., Grenier, J.K., and Weatherbee, S.D. (2013). From DNA to diversity: molecular genetics and the evolution of animal design (John Wiley & Sons).

19. Clelland, B.W., and Schultz, M.C. (2010). Genome stability control by checkpoint regulation of tRNA gene transcription. Transcription *1*, 115-125.

20. Conesa, C., Ruotolo, R., Soularue, P., Simms, T.A., Donze, D., Sentenac, A., and Dieci, G. (2005). Modulation of yeast genome expression in response to defective RNA polymerase III-dependent transcription. Mol Cell Biol *25*, 8631-8642.

21. Cordaux, R., and Batzer, M.A. (2009). The impact of retrotransposons on human genome evolution. Nat Rev Genet *10*, 691-703.

22. Crick, F. (1970). Central dogma of molecular biology. Nature *227*, 561-563.

23. Crick, F.H. (1958). On protein synthesis. Symposia of the Society for Experimental Biology *12*, 138-163.

24. D'Ambrosio, C., Schmidt, C.K., Katou, Y., Kelly, G., Itoh, T., Shirahige, K., and Uhlmann, F. (2008). Identification of cis-acting sites for condensin loading onto budding yeast chromosomes. Genes Dev *22*, 2215-2227.

25. Davidson, E.H., and Erwin, D.H. (2006). Gene regulatory networks and the evolution of animal body plans. Science *311*, 796-800.

26. Degnan, B.M., Vervoort, M., Larroux, C., and Richards, G.S. (2009). Early evolution of metazoan transcription factors. Current opinion in genetics & development *19*, 591-599.

27. Dieci, G., Fiorino, G., Castelnuovo, M., Teichmann, M., and Pagano, A. (2007). The expanding RNA polymerase III transcriptome. TRENDS in Genetics *23*, 614-622.

28. Dittmar, K.A., Goodenbour, J.M., and Pan, T. (2006). Tissue-specific differences in human transfer RNA expression. PLoS Genet *2*, e221.

29. Doebley, J., and Lukens, L. (1998). Transcriptional regulators and the evolution of plant form. The Plant Cell *10*, 1075-1082.

30. Donze, D., Adams, C.R., Rine, J., and Kamakaka, R.T. (1999). The boundaries of the silenced HMR domain in Saccharomyces cerevisiae. Genes & development *13*, 698-708.

31. Donze, D., and Kamakaka, R.T. (2001). RNA polymerase III and RNA polymerase II promoter complexes are heterochromatin barriers in Saccharomyces cerevisiae. The EMBO journal *20*, 520-531.

32. Ebersole, T., Kim, J.-H., Samoshkin, A., Kouprina, N., Pavlicek, A., White, R.J., and Larionov, V. (2011). tRNA genes protect a reporter gene from epigenetic silencing in mouse cells. Cell Cycle *10*, 2779-2791.

33. Englander, E.W., Wolffe, A., and Howard, B.H. (1993). Nucleosome interactions with a human Alu element. Transcriptional repression and effects of template methylation. Journal of Biological Chemistry *268*, 19565-19573.

34. Filippova, G.N., Cheng, M.K., Moore, J.M., Truong, J.-P., Hu, Y.J., Tsuchiya, K.D., and Disteche, C.M. (2005). Boundaries between chromosomal domains of X inactivation and escape bind CTCF and lack CpG methylation during early development. Developmental cell *8*, 31-42.

35. Flinn, E.M., Wallberg, A.E., Hermann, S., Grant, P.A., Workman, J.L., and Wright, A.P. (2002). Recruitment of Gcn5-containing Complexes during c-Myc-dependent Gene Activation STRUCTURE AND FUNCTION ASPECTS. Journal of Biological Chemistry *277*, 23399-23406.

36. Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.-l., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. Genetics *151*, 1531-1545.

37. Frankel, A.D., Berg, J.M., and Pabo, C.O. (1987). Metal-dependent folding of a single zinc finger from transcription factor IIIA. Proceedings of the National Academy of Sciences *84*, 4841-4845.

38. Geiduschek, E.P., and Kassavetis, G.A. (2001). The RNA polymerase III transcription apparatus. Journal of molecular biology *310*, 1-26.

39. Gelbart, M.E., Bachman, N., Delrow, J., Boeke, J.D., and Tsukiyama, T. (2005). Genome-wide identification of Isw2 chromatin-remodeling targets by localization of a catalytically inactive mutant. Genes & development *19*, 942-954.

40. Gilad, Y., Rifkin, S.A., and Pritchard, J.K. (2008). Revealing the architecture of gene regulation: the promise of eQTL studies. Trends in genetics *24*, 408-415.

41. Gomez-Roman, N., Grandori, C., Eisenman, R.N., and White, R.J. (2003). Direct activation of RNA polymerase III transcription by c-Myc. Nature *421*, 290-294.

42. Grove, C.A., De Masi, F., Barrasa, M.I., Newburger, D.E., Alkema, M.J., Bulyk, M.L., and Walhout, A.J. (2009). A multiparameter network reveals extensive divergence between C. elegans bHLH transcription factors. Cell *138*, 314-327.

43. Haeusler, R.A., Pratt-Hyatt, M., Good, P.D., Gipson, T.A., and Engelke, D.R. (2008). Clustering of yeast tRNA genes is mediated by specific association of condensin with tRNA gene transcription complexes. Genes & development *22*, 2204-2214.

44. Hakimi, M.-A., Bochar, D.A., Schmiesing, J.A., Dong, Y., Barak, O.G., Speicher, D.W., Yokomori, K., and Shiekhattar, R. (2002). A chromatin remodelling complex that loads cohesin onto human chromosomes. Nature *418*, 994-998.

45. Harismendy, O., Gendrel, C.G., Soularue, P., Gidrol, X., Sentenac, A., Werner, M., and Lefebvre, O. (2003). Genome-wide location of yeast RNA polymerase III transcription machinery. The EMBO Journal *22*, 4738-4747.

46. Hirata, T., Amano, T., Nakatake, Y., Amano, M., Piao, Y., Hoang, H.G., and Ko, M.S. (2012). Zscan4 transiently reactivates early embryonic genes during the generation of induced pluripotent stem cells. Scientific reports *2*.

47. Hoekstra, H.E., and Coyne, J.A. (2007). The locus of evolution: evo devo and the genetics of adaptation. Evolution *61*, 995-1016.

48. Hsia, C.C., and McGinnis, W. (2003). Evolution of transcription factor function. Current opinion in genetics & development *13*, 199-206.

49. Huang, Y., and Maraia, R.J. (2001). Comparison of the RNA polymerase III transcription machinery in Schizosaccharomyces pombe, Saccharomyces cerevisiae and human. Nucleic Acids Research *29*, 2675-2690.

50. Hull, M.W., Erickson, J., Johnston, M., and Engelke, D.R. (1994). tRNA genes as transcriptional repressor elements. Molecular and cellular biology *14*, 1266-1277.

51. Huntley, S., Baggott, D.M., Hamilton, A.T., Tran-Gyamfi, M., Yang, S., Kim, J., Gordon, L., Branscomb, E., and Stubbs, L. (2006). A comprehensive catalog of human KRAB-associated zinc finger genes: insights into the evolutionary history of a large family of transcriptional repressors. Genome research *16*, 669-677.

52. Iuchi, S. (2001). Three classes of C2H2 zinc finger proteins. Cellular and Molecular Life Sciences CMLS *58*, 625-635.

53. Jacob, F., and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. Journal of molecular biology *3*, 318-356.

54. Jin, C., Zang, C., Wei, G., Cui, K., Peng, W., Zhao, K., and Felsenfeld, G. (2009). H3. 3/H2A. Z double variant–containing nucleosomes mark'nucleosome-free regions' of active promoters and other regulatory regions. Nature genetics *41*, 941-945.

55. Jurka, J., Zietkiewicz, E., and Labuda, D. (1995). Ubiquitous mammalian-wide interspersed repeats (MIRs) are molecular fossils from the mesozoic era. Nucleic Acids Research *23*, 170-175.

56. Karin, M. (1990). Too many transcription factors: positive and negative interactions. The New Biologist *2*, 126-131.

57. Kendall, A., Hull, M.W., Bertrand, E., Good, P.D., Singer, R.H., and Engelke, D.R. (2000). A CBF5 mutation that disrupts nucleolar localization of early tRNA biosynthesis in yeast also suppresses tRNA gene-mediated transcriptional silencing. Proceedings of the National Academy of Sciences *97*, 13108-13113.

58. Kenneth, N.S., Ramsbottom, B.A., Gomez-Roman, N., Marshall, L., Cole, P.A., and White, R.J. (2007). TRRAP and GCN5 are used by c-Myc to activate RNA polymerase III transcription. Proceedings of the National Academy of Sciences *104*, 14917-14922.

59. Kim, C., Rubin, C.M., and Schmid, C.W. (2001). Genome-wide chromatin remodeling modulates the Alu heat shock response. Gene *276*, 127-133.

60. Kinsey, P.T., and Sandmeyer, S.B. (1991). Adjacent pol II and pol III promoters: transcription of the yeast retrotransposon Ty3 and a target tRNA gene. Nucleic acids research *19*, 1317-1324.

61. Klug, A. (2010). The discovery of zinc fingers and their applications in gene regulation and genome manipulation. Annual review of biochemistry *79*, 213-231.

62. Klug, A., and Schwabe, J. (1995). Protein motifs 5. Zinc fingers. The FASEB journal *9*, 597-604.

63. Kraus, P., Sivakamasundari, V., Yu, H.B., Xing, X., Lim, S.L., Adler, T., Pimentel, J.A.A., Becker, L., Bohla, A., and Garrett, L. (2014). Pleiotropic functions for transcription factor zscan10. PloS one *9*, e104568.

64. Lamb, P., and McKnight, S.L. (1991). Diversity and specificity in transcriptional regulation: the benefits of heterotypic dimerization. Trends in biochemical sciences *16*, 417-422.

65. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., and FitzHugh, W. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

66. Latchman, D.S. (1997). Transcription factors: an overview. The international journal of biochemistry & cell biology *29*, 1305-1312.

67. Lee, T.I., and Young, R.A. (2000). Transcription of eukaryotic protein-coding genes. Annual review of genetics *34*, 77-137.

68. Levine, M., and Tjian, R. (2003). Transcription regulation and animal diversity. Nature *424*, 147-151.

69. Li, T.-H., Kim, C., Rubin, C.M., and Schmid, C.W. (2000). K562 cells implicate increased chromatin accessibility in Alu transcriptional activation. Nucleic acids research *28*, 3031-3039.

70. Liu, H., Chang, L.-H., Sun, Y., Lu, X., and Stubbs, L. (2014). Deep vertebrate roots for mammalian zinc finger transcription factor subfamilies. Genome biology and evolution *6*, 510-525.

71. Liu, W.-M., Maraia, R.J., Rubin, C.M., and Schmid, C.W. (1994). Alu transcripts: cytoplasmic localisation and regulation by DNA methylation. Nucleic acids research *22*, 1087-1095.

72. Liu, W.-M., and Schmid, C.W. (1993). Proposed roles for DNA methylation in Alu transcriptional repression and mutational inactivation. Nucleic acids research *21*, 1351-1359.

73. Lunyak, V.V., Prefontaine, G.G., Núñez, E., Cramer, T., Ju, B.-G., Ohgi, K.A., Hutt, K., Roy, R., García-Díaz, A., and Zhu, X. (2007). Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis. Science *317*, 248-251.

74. Lynch, V.J., May, G., and Wagner, G.P. (2011). Regulatory evolution through divergence of a phosphoswitch in the transcription factor CEBPB. Nature *480*, 383-386.

75. Magdinier, F., Yusufzai, T.M., and Felsenfeld, G. (2004). Both CTCF-dependent and-independent insulators are found between the mouse T cell receptor α and Dad1 genes. Journal of Biological Chemistry *279*, 25381-25389.

76. Mishiro, T., Ishihara, K., Hino, S., Tsutsumi, S., Aburatani, H., Shirahige, K., Kinoshita, Y., and Nakao, M. (2009). Architectural roles of multiple chromatin insulators at the human apolipoprotein gene cluster. The EMBO journal *28*, 1234-1245.

77. Moqtaderi, Z., and Struhl, K. (2004). Genome-wide occupancy profile of the RNA polymerase III machinery in Saccharomyces cerevisiae reveals loci with incomplete transcription complexes. Molecular and cellular biology *24*, 4118-4127.

78. Moqtaderi, Z., Wang, J., Raha, D., White, R.J., Snyder, M., Weng, Z., and Struhl, K. (2010). Genomic binding profiles of functionally distinct RNA polymerase III transcription complexes in human cells. Nature structural & molecular biology *17*, 635-640.

79. Nakagawa, M., Takizawa, N., Narita, M., Ichisaka, T., and Yamanaka, S. (2010). Promotion of direct reprogramming by transformation-deficient Myc. Proceedings of the National Academy of Sciences *107*, 14152-14157.

80. Nikolov, D., and Burley, S. (1997). RNA polymerase II transcription initiation: a structural view. Proceedings of the National Academy of Sciences *94*, 15-22.

81. Noma, K.-i., Cam, H.P., Maraia, R.J., and Grewal, S.I. (2006). A role for TFIIIC transcription factor complex in genome organization. Cell *125*, 859-872.

82. Nowick, K., Fields, C., Gernat, T., Caetano-Anolles, D., Kholina, N., and Stubbs, L. (2011). Gain, loss and divergence in primate zinc-finger genes: a rich resource for evolution of gene regulatory differences between species. PLoS One *6*, e21553.

83. Nowick, K., and Stubbs, L. (2010). Lineage-specific transcription factors and the evolution of gene regulatory networks. Briefings in functional genomics, elp056.

84. Odom, D.T., Dowell, R.D., Jacobsen, E.S., Gordon, W., Danford, T.W., MacIsaac, K.D., Rolfe, P.A., Conboy, C.M., Gifford, D.K., and Fraenkel, E. (2007). Tissue-specific transcriptional regulation has diverged significantly between human and mouse. Nature genetics *39*, 730-732.

85. Oler, A.J., Alla, R.K., Roberts, D.N., Wong, A., Hollenhorst, P.C., Chandler, K.J., Cassiday, P.A., Nelson, C.A., Hagedorn, C.H., and Graves, B.J. (2010). Human RNA polymerase III transcriptomes and relationships to Pol II promoter chromatin and enhancer-binding factors. Nature structural & molecular biology *17*, 620-628.

86. Orioli, A., Pascali, C., Pagano, A., Teichmann, M., and Dieci, G. (2012). RNA polymerase III transcription control elements: themes and variations. Gene *493*, 185-194.

87. Parelho, V., Hadjur, S., Spivakov, M., Leleu, M., Sauer, S., Gregson, H.C., Jarmuz, A., Canzonetta, C., Webster, Z., and Nesterova, T. (2008). Cohesins functionally associate with CTCF on mammalian chromosome arms. Cell *132*, 422-433.

88. Pascali, C., and Teichmann, M. (2013). RNA Polymerase III Transcription–Regulated by Chromatin Structure and Regulator of Nuclear Chromatin Organization. In Epigenetics: Development and Disease (Springer), pp. 261-287.

89. Paule, M.R., and White, R.J. (2000). Survey and summary transcription by RNA polymerases I and III. Nucleic acids research *28*, 1283-1298.

90. Pavletich, N.P., and Pabo, C.O. (1991). Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. Science *252*, 809-817.

91. Pengue, G., Calabró, V., Bartoli, P.C., Pagliuca, A., and Lania, L. (1994). Repression of transcriptional activity at a distance by the evolutionary conserved KRAB domain present in a subfamily of zinc finger proteins. Nucleic Acids Research *22*, 2908-2914.

92. Pennacchio, L.A., Bickmore, W., Dean, A., Nobrega, M.A., and Bejerano, G. (2013). Enhancers: five essential questions. Nature Reviews Genetics *14*, 288-295.

93. Pérez, J.C., Fordyce, P.M., Lohse, M.B., Hanson-Smith, V., DeRisi, J.L., and Johnson, A.D. (2014). How duplicated transcription regulators can diversify to govern the expression of nonoverlapping sets of genes. Genes & development *28*, 1272-1277.

94. Phillips, J.E., and Corces, V.G. (2009). CTCF: master weaver of the genome. Cell *137*, 1194-1211.

95. Prud'homme, B., Gompel, N., and Carroll, S.B. (2007). Emerging principles of regulatory evolution. Proceedings of the National Academy of Sciences *104*, 8605-8612.

96. Raab, J.R., Chiu, J., Zhu, J., Katzman, S., Kurukuti, S., Wade, P.A., Haussler, D., and Kamakaka, R.T. (2012). Human tRNA genes function as chromatin insulators. The EMBO journal *31*, 330-350.

97. Reece-Hoyes, J.S., Pons, C., Diallo, A., Mori, A., Shrestha, S., Kadreppa, S., Nelson, J., DiPrima, S., Dricot, A., and Lajoie, B.R. (2013). Extensive rewiring and complex evolutionary dynamics in a C. elegans multiparameter transcription factor network. Molecular cell *51*, 116-127.

98. Roberts, D.N., Stewart, A.J., Huff, J.T., and Cairns, B.R. (2003). The RNA polymerase III transcriptome revealed by genome-wide localization and activity–

occupancy relationships. Proceedings of the National Academy of Sciences *100*, 14695-14700.

99. Roeder, R.G. (1996). The role of general initiation factors in transcription by RNA polymerase II. Trends in biochemical sciences *21*, 327-335.

100. Román, A.C., González-Rico, F.J., Moltó, E., Hernando, H., Neto, A., Vicente-Garcia, C., Ballestar, E., Gómez-Skarmeta, J.L., Vavrova-Anderson, J., and White, R.J. (2011). Dioxin receptor and SLUG transcription factors regulate the insulator activity of B1 SINE retrotransposons via an RNA polymerase switch. Genome research *21*, 422-432.

101. Rubinstein, M., and de Souza, F.S. (2013). Evolution of transcriptional enhancers and animal diversity. Philosophical Transactions of the Royal Society of London B: Biological Sciences *368*, 20130017.

102. Russanova, V.R., Driscoll, C.T., and Howard, B.H. (1995). Adenovirus type 2 preferentially stimulates polymerase III transcription of Alu elements by relieving repression: a potential role for chromatin. Molecular and cellular biology *15*, 4282-4290.

103. Sander, T.L., Haas, A.L., Peterson, M.J., and Morris, J.F. (2000). Identification of a Novel SCAN Box-related Protein That Interacts with MZF1B the Leucine-Rich SCAN box Mediates Hetero- and Homo-protein Associations. Journal of Biological Chemistry *275*, 12857-12867.

104. Sander, T.L., Stringer, K.F., Maki, J.L., Szauter, P., Stone, J.R., and Collins, T. (2003). The SCAN domain defines a large family of zinc finger transcription factors. Gene *310*, 29-38.

105. Sayou, C., Monniaux, M., Nanao, M.H., Moyroud, E., Brockington, S.F., Thévenon, E., Chahtane, H., Warthmann, N., Melkonian, M., and Zhang, Y. (2014). A promiscuous intermediate underlies the evolution of LEAFY DNA binding specificity. Science *343*, 645-648.

106. Schumacher, C., Wang, H., Honer, C., Ding, W., Koehn, J., Lawrence, Q., Coulis, C.M., Wang, L.L., Ballinger, D., and Bowen, B.R. (2000). The SCAN domain mediates selective oligomerization. Journal of Biological Chemistry *275*, 17173-17179.

107. Shaikh, T.H., Roy, A.M., Kim, J., Batzer, M.A., and Deininger, P.L. (1997). cDNAs derived from primary and small cytoplasmic Alu (scAlu) transcripts. Journal of molecular biology *271*, 222-234.

108. Silva, J., Shabalina, S., Harris, D., Spouge, J., and Kondrashov, A. (2003). Conserved fragments of transposable elements in intergenic regions: evidence for widespread recruitment of MIR-and L2-derived sequences within the mouse and human genomes. Genetical research *82*, 1-18.

109. Simms, T.A., Dugas, S.L., Gremillion, J.C., Ibos, M.E., Dandurand, M.N., Toliver, T.T., Edwards, D.J., and Donze, D. (2008). TFIIIC binding sites function as both heterochromatin barriers and chromatin insulators in Saccharomyces cerevisiae. Eukaryotic cell *7*, 2078-2086.

110. Smale, S.T., and Kadonaga, J.T. (2003). The RNA polymerase II core promoter. Annual review of biochemistry *72*, 449-479.

111. Smit, A.F., and Riggs, A.D. (1995). MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. Nucleic Acids Research *23*, 98-102.

112. Stone, J.R., Maki, J.L., Blacklow, S.C., and Collins, T. (2002). The SCAN domain of ZNF174 is a dimer. Journal of Biological Chemistry *277*, 5448-5452.

113. Stubbs, L., Sun, Y., and Caetano-Anolles, D. (2011). Function and evolution of C2H2 zinc finger arrays. In A Handbook of Transcription Factors (Springer), pp. 75-94.

114. Teichmann, S.A., and Babu, M.M. (2004). Gene regulatory network growth by duplication. Nature genetics *36*, 492-496.

115. Thompson, M., Haeusler, R.A., Good, P.D., and Engelke, D.R. (2003). Nucleolar clustering of dispersed tRNA genes. Science *302*, 1399-1401.

116. Urrutia, R. (2003). KRAB-containing zinc-finger repressor proteins. Genome Biol *4*, 231.

117. Van Bortle, K., and Corces, V.G. (2012). tDNA insulators and the emerging role of TFIIIC in genome organization. Transcription *3*, 277-284.

118. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., and Holt, R.A. (2001). The sequence of the human genome. science *291*, 1304-1351.

119. Wang, L., Haeusler, R.A., Good, P.D., Thompson, M., Nagar, S., and Engelke, D.R. (2005). Silencing near tRNA genes requires nucleolar localization. Journal of Biological Chemistry *280*, 8637-8639.

120. Wang, Z.-X., Teh, C.H.-L., Kueh, J.L., Lufkin, T., Robson, P., and Stanton, L.W. (2007). Oct4 and Sox2 directly regulate expression of another pluripotency transcription factor, Zfp206, in embryonic stem cells. Journal of Biological Chemistry *282*, 12822-12830.

121. Wendt, K.S., Yoshida, K., Itoh, T., Bando, M., Koch, B., Schirghuber, E., Tsutsumi, S., Nagae, G., Ishihara, K., and Mishiro, T. (2008). Cohesin mediates transcriptional insulation by CCCTC-binding factor. Nature *451*, 796-801.

122. Williams, A.J., Blacklow, S.C., and Collins, T. (1999). The zinc finger-associated SCAN box is a conserved oligomerization domain. Molecular and cellular biology *19*, 8526-8535.

123. Williams, A.J., Khachigian, L.M., Shows, T., and Collins, T. (1995). Isolation and characterization of a novel zinc-finger protein with transcriptional repressor activity. Journal of Biological Chemistry *270*, 22143-22152.

124. Wittkopp, P. (2005). Genomic sources of regulatory variation in cis and in trans. Cellular and Molecular Life Sciences CMLS *62*, 1779-1783.

125. Wolfe, S.A., Nekludova, L., and Pabo, C.O. (2000). DNA recognition by Cys2His2 zinc finger proteins. Annual review of biophysics and biomolecular structure *29*, 183-212.

126. Wood, A.J., Severson, A.F., and Meyer, B.J. (2010). Condensin and cohesin complexity: the expanding repertoire of functions. Nature Reviews Genetics *11*, 391-404.

127. Wray, G.A. (2007). The evolutionary significance of cis-regulatory mutations. Nature Reviews Genetics *8*, 206-216.

128. Yvert, G., Brem, R.B., Whittle, J., Akey, J.M., Foss, E., Smith, E.N., Mackelprang, R., and Kruglyak, L. (2003). Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. Nature genetics *35*, 57-64.

CHAPTER 2:

ZSCAN5B AND ITS PRIMATE-SPECIFIC PARALOGS BIND RNA POLYMERASE III GENES AND EXTRA-TFIIIC (ETC) SITES IN MAMMALIAN CELLS

Younguk Sun[1,2], Huimin Zhang[1,2], Majid Kazemian[1,3,6], Joseph M. Troy[1,4], Christopher Seward[1,2], Xiaochen Lu[1,2], Saurabh Sinha[1,3], and Lisa Stubbs[1,2,4,5]

ABSTRACT

Particularly in the context of differentiation and development, the importance of three-dimensional chromatin architecture on gene regulation is becoming increasingly clear. The most ancient known mechanism of chromatin organization involves TFIIIC, a transcription factor (TF) that recruits RNA polymerase III (Pol III) for transcription of tRNA and other types of non-coding RNA genes. From yeast to mammals, TFIIIC also binds to scattered "extra-TFIIIC" (ETC) loci, serving to tether those loci together as anchors of chromatin loops. TFIIIC activities are modulated by MAF, MYC, and other TF proteins, many of which are likely unknown. Here we identify the ZSCAN5 TF family - including mammalian ZSCAN5B and its primate-specific paralogs - as proteins that occupy mammalian Pol III promoters and ETC sites. ZSCAN5B binds with high specificity to a subset of conserved tRNA genes (tDNA) in human and mouse chromatin, and modulates the expression of those and other Pol III genes. Primate-specific ZSCAN5A and ZSCAN5D also bind tDNA, although ZSCAN5D preferentially localizes to MIR SINE- and LINE2-associated ETC sites. *ZSCAN5A* and *ZSCAN5B* knockdown experiments indicate important and cooperative roles in regulating tRNA and rRNA processing, cell cycle progression and differentiation in a variety of tissues. Together these data indicate that ZSCAN5 protein binding directly regulates levels of Pol III gene transcription, thereby exerting position effects on nearby Pol II genes; ultimately, this activity influences cell cycle progression and differentiation during development and in adult tissues.

[1] Institute for Genomic Biology

[2] Department of Cell and Developmental Biology

[3] Department of Computer Science

[4] Illinois Informatics Program,

University of Illinois at Urbana-Champaign, Urbana IL 61801

[5] Corresponding author

Keywords:

Zinc finger transcription factor/Primate-specific duplication/RNA Polymerase III transcription/tRNA/chromatin architecture

A modified version of this manuscript was submitted to Oncotarget: Chromosome (in review).

INTRODUCTION

In traditional models of gene regulation, transcription factors (TFs) interact locally to enhance or block RNA polymerase recruitment to nearby promoters. However, long-range interactions enabled by chromatin looping are increasingly being recognized for essential regulatory roles (Levine et al., 2014). The most ancient known mechanism of chromatin loop organization is dependent on TFIIIC, a deeply conserved TF that recruits TFIIIB, TFIIIA and RNA polymerase III (Pol III) to promoters of tRNA genes (tDNAs) and other types of Pol III transcripts (Pascali and Teichmann, 2013). Through association with TFIIIC, tDNAs and other binding sites cluster within the nucleus to serve as key determinants of three-dimensional genomic architecture and exert chromatin barrier, insulator and other regulatory functions. These functions appear to be remarkably well conserved across eukarya, as confirmed by studies in yeast, insects, and mammals (Crepaldi et al., 2013; Hiraga et al., 2012; Van Bortle et al., 2014; Wang et al., 2014).

Most vertebrate tDNAs are organized in genomic clusters at homologous, syntenic positions, thus providing a stable and conserved framework for chromatin structure (Van Bortle and Corces, 2012). However not all TFIIIC binding sites are so highly conserved. In particular, extra-TFIIIC, or ETC sites, also called “chromatin organizing clamps” in yeast, interact with each other and with tDNA to influence chromatin architecture (Noma et al., 2006; Raab et al., 2012). In mammals, Pol III and TFIIIC binding sites include SINE repeats, transposable elements including MIRs, ALUs and other SINE subfamilies that were originally derived from Pol III transcription units (Cordaux and Batzer, 2009). These lineage-specific transposable elements (TEs) greatly outnumber the evolutionarily more stable tDNA sites in mammalian genomes.

Furthermore, the expression of individual tRNA genes varies according to cell type, cellular states, tissues and developmental stages (Canella et al., 2012; Canella et al., 2010; Dittmar et al., 2006; Schmitt et al., 2014; Stutz et al., 1989). With or without transcription, TFIIIC binding also varies at tDNAs and at ETC sites (Barski et al., 2010; Canella et al., 2010; Moqtaderi et al., 2010; Raha et al., 2010; Schmitt et al., 2014). TFIIIC, TFIIIB and Pol III recruitment steps are modulated by interactions with transcription factors including MAF1, MYC, RB1, P53, and very likely, many other

unknown TF proteins (Felton-Edkins et al., 2003; Hiraga et al., 2012; Oler et al., 2010; Raha et al., 2010). The influence of these TFs on tRNA expression has direct developmental consequences (Rideout et al., 2012) as well as indirect effects on transcription of nearby Pol II genes (Kinsey and Sandmeyer, 1991; Lee et al., 2015; Moqtaderi et al., 2010; Wang et al., 2014). TFIIIC recruitment to tDNAs and ETCs also influences nuclear clustering and thus likely, the selection of alternative anchors for chromatin loops (Kirkland et al., 2013; Lee et al., 2015; Raab et al., 2012; Van Bortle and Corces, 2012).

Here we report the DNA-binding functions of ZSCAN5B, a protein encoded by a SCAN domain-containing zinc finger (SCAN-ZNF) gene that is unique in most eutherian species but was duplicated to generate three paralogs in early primate lineages, two of which we also investigate in this study. The paralogs have diverged from *ZSCAN5B* in both zinc-finger structure and expression patterns but overlap in several types of cells and tissues. Chromatin immunoprecipitation followed by deep sequencing (ChIP-seq) revealed that tDNA sequences are the strongly preferred binding sites for mouse and human ZSCAN5B and also for human ZSCAN5A; paralog ZSCAN5D also bind tDNAs, but favors binding to MIR SINE and LINE2 elements including a previously identified ETC-related motif (Moqtaderi et al., 2010). *ZSCAN5A* and *ZSCAN5B* gene knockdown resulted in higher expression of bound Pol III transcripts and dysregulation of nearby polymerase II (Pol II)-expressed genes, enriched in functions including the regulation of tRNA and rRNA processing, cell-fate decisions, and control of mitosis. Based on these data, we conjecture that *ZSCAN5B* evolved in eutherians to directly modulate the expression and interaction of bound Pol III transcription units, and that *ZSCAN5A* and *ZSCAN5D* evolved in primates to extend this negative regulatory activity to include a wider range of TFIIIC binding sites, including MIR and L2 repeat-associated ETCs.

RESULTS

The ZSCAN5 family arose by duplication of conserved *Zscan5b* in early primate history

Zscan5b is a unique gene in mouse and most other eutherian genomes, but primate genomes contain four very closely related gene copies, annotated as human *ZSCAN5A, ZSCAN5B, ZSCAN5C* and *ZSCAN5D* (Harrow et al., 2006). Human *ZSCAN5B* is clearly the ortholog of the single mouse gene as confirmed by overall sequence similarity as well as the alignment of the DNA-binding amino acids of each zinc finger (corresponding to amino acids -1, 2, 3, and 6 relative to the alpha-helix) (Berg, 1992; Pavletich and Pabo, 1991; Wolfe et al., 2001) (*Fig. 2.1*). For simplicity, as we have in a recent paper (Liu et al., 2014), we will refer to this pattern of DNA-binding amino acid quadruplets as a protein's "fingerprint" henceforth.

Mouse, marmoset and human ZSCAN5B proteins share almost identical fingerprints, but the three primate-specific paralogs have diverged from the parental gene in fingerprint patterns. After an initial phase of divergence in early primate history, these patterns have been very well conserved (*Fig. 2.1***)**. In contrast to the zinc fingers, the SCAN domains of the four human ZSCAN5 family members are nearly identical (95-98% identity between members; not shown). Since SCAN mediates protein dimer formation (Collins et al., 2001; Schumacher et al., 2000), this suggests that ZSCAN5 family members should interact as heterodimers, cooperating in different combinatorial patterns in tissues and cell types where they are co-expressed.

We found orthologs of all four human genes in the marmoset genome but identified only a unique *ZSCAN5B*-related copy in other eutherian genomes and in genomes of more primitive primates such as *Galago* (not shown). Available evidence therefore indicates that *ZSCAN5A, ZSCAN5C,* and *ZSCAN5D* genes are present only in new world monkeys and later-evolving primates.

ZSCAN5 paralogs display overlapping but unique patterns of expression

Overlapping but distinct patterns of tissue-specific expression in human tissues

Publicly available data indicated that both human and mouse ZSCAN5-family genes are expressed at high levels in testis but at very low levels in most other adult tissues. To further examine expression profiles of ZSCAN5 family members, we used quantitative reverse transcript PCR (qRT-PCR) to measure transcript levels of the unique mouse *Zscan5b* and all four human ZSCAN5 genes in panels of RNA derived from adult and embryonic tissues (*Fig. 2.2A*). *ZSCAN5C* transcripts were not detected, or were detected near background levels, in every tissue we tested (not shown). As expected, mouse *Zscan5b* and all three expressed human paralogs were detected at highest levels in adult testis. In adult mouse, we found the second highest levels of *Zscan5b* transcript in thymus, fetal liver and placenta with lower levels of expression in brain, lung, and skeletal muscle. Across a similar panel of tissues, the human ZSCAN5 paralogs were expressed in patterns that differed from that of the mouse gene; *ZSCAN5A* and *ZSCAN5B* were similarly expressed whereas *ZSCAN5D* was the most distinct. *ZSCAN5A* transcripts were detected robustly in most tissues, and at much higher absolute levels than any of the other primate paralogs. Therefore, the human duplicates have acquired distinct features of gene expression – either in terms of tissue-specific patterns or overall levels of expression – compared to each other and compared to the unique mouse gene.

Mouse and human ZSCAN5B are expressed in populations of actively dividing cells

To identify cell type-specific expression patterns for the conserved paralog *in vivo,* we developed probes for *in situ* hybridization (ISH) from the unique 3'-untranslated (3'UTR) regions of the mouse and human *ZSCAN5B* genes. For mouse, we hybridized probes to sagittal section of whole embryos taken at 14.5 days post-coitum (E14.5), E16.5 and E18.5; for human, we examined paraffin sections of a selection of adult tissues on a tissue array. Mouse *Zscan5b* displayed highest expression in E14.5 heart (*Fig. 2.3A, 2.3C*), alveoli of the developing lungs, spinal cord and forebrain (*Fig. 2.3A, 2.3B*). Heart and lung expression was diminished but expression remained particularly high in the

olfactory bulb (*Fig. 2.3D*) and thymus (*Fig. 2.3E*) at E16.5. By E18.5, expression was high in developing skeletal muscle and skin (*Fig. 2.3F*) and cartilage and lower level in bone in the vertebral column (*Fig. 2.3G*), and in forebrain as well as the intestinal epithelia (not shown). In human tissues, *ZSCAN5B* was also detected in adult skin (*Fig. 2.3H*), epithelial cells in the small intestine (*Fig. 2.3I*), testicular spermatocytes (*Fig. 2.3J*), lung epithelia (*Fig. 2.3K*), and bone marrow (*Fig. 2.3L*); thymocytes were also strongly positive for human *ZSCAN5B*, while tissue cores taken from several adult brain regions were not (not shown). In general, human adult gene expression was highest in tissues and cell types with high levels of cellular turnover and cell division. Interestingly, despite qPCR data suggesting different patterns of overall tissue-specific expression (*Fig. 2.2A*), the cell types and tissues that express *Zscan5b* in mouse embryos overlap considerably with those that display high levels of human *ZSCAN5B* in adults.

Gene knockdown experiments reveal clues to shared and unique cellular functions

Testing RNA from a panel of human cell lines with qRT-PCR confirmed the high and virtually ubiquitous expression of ZSCAN5A in cultured cells. However, very few cell lines also expressed ZSCAN5B or ZSCAN5D (not shown). With the goal of examining paralog function in the same cellular context, and we identified two lines – BeWo, a trophoblast-like cell line derived from choriocarcinoma, and HEK-293, derived from human embryonic kidney but with neuronal characteristics (Shaw et al., 2002) – in which all three paralogs were expressed, and focused further on experiments with those cell lines.

Since transfection is particularly efficient for HEK-293 cells, we used this cell line for siRNA knockdown experiments. We tested a number of independent siRNA designs, both commercially available and custom, for each gene to assess efficiency and specificity of paralog knockdown. Most siRNA designs displayed off-target effects significantly affecting at least two of the ZSCAN5 genes under conditions we tested (not shown). However, two of the siRNA reagents (hereafter called si4 and si5) knocked down levels of *ZSCAN5A* relatively efficiently, with lesser effect on *ZSCAN5D* but marginal levels of knockdown (si5) or even over-expression (si4) for *ZSCAN5B.* Additionally, we found one siRNA design that knocked down *ZSCAN5B* transcripts quite

well and specifically (si1). We identified one siRNA design (si2) that allowed a reasonable degree of *ZSCAN5D* knockdown without reducing levels of either of the other two ZSCAN5 genes, but treatment with si2 increased *ZSCAN5B* transcripts levels significantly in HEK-293 cells (*Fig. 2.4A, Table A.1*). Unlike *ZSCAN5A,* we could not find a second siRNA that knocked down *ZSCAN5D* without significantly effecting the levels of other genes, and this issue complicated further functional analysis for this paralog.

We analyzed RNA from cells treated with si1, si4, si5, and a scrambled control using RNA-seq to identify differentially expressed genes (DEGs). Since two siRNA designs could be tested independently, the *ZSCAN5A* siRNA analysis yielded the most robust and reliable DEG set (363 genes detected with adjusted $P \geq 0.05$, fold change $\geq$ 1.5; *Table A.2*). Analyzing these DEGs with the DAVID functional analysis program (Huang *et al.* 2009) revealed very high enrichment for specific functional categories in the up- and down-regulated genes (*Table 2.1*). Notably, genes up-regulated after *ZSCAN5A* knockdown (e.g. negatively regulated by ZSCAN5A) included regulators of ribosome biogenesis and cell cycle regulators especially proteins controlling spindle attachment, chromosome segregation and the metaphase/anaphase transition. Down-regulated DEGs (positively regulated by ZSCAN5A) were significantly enriched for functions including transcriptional regulation, cell adhesion, and cell-fate commitment in a variety of tissues.

Genes differentially expressed after human *ZSCAN5B* knockdown displayed significant levels of overlap with *ZSCAN5A* DEGs, with 100 of the 363 *ZSCAN5A* DEGs being detected as similarly up- or down-regulated in the *ZSCAN5B* siRNA experiment (*Table A.2*). Since the *ZSCAN5A* knockdown experiments we analyzed did not reduce levels of *ZSCAN5B* and vice versa (*Fig. 2.4*), these data suggested some level of functional cooperation between the paralogous proteins in HEK-293 cells. Nevertheless the *ZSCAN5B* DEG set also emphasized some novel functions especially those related to tRNA and rRNA processing and modification (up-regulated) and microRNA processing (down-regulated) (*Table 2.1*). Interestingly, the tissues predicted to be affected by ZSCAN5 functions –including kidney, gut, cartilage, hematopoietic/lymphoid tissues,

and olfactory bulb - overlapped well with expression sites determined by qRT-PCR and ISH for the human and/or mouse genes (*Table 2.1*; *Fig. 2.2*, *Fig. 2.3*).

ZSCAN5 proteins bind to tRNA genes and extra-TFIIIC sites in human and mouse

ChIP with paralog-specific antibodies

To identify reagents for detection of ZSCAN5 proteins, we identified commercial antibodies targeting ZSCAN5B and ZSCAN5D, and designed peptide epitopes from a sequence-divergent region of human ZSCAN5A and from the mouse Zscan5b protein to generate custom polyclonal antibodies (see Methods). These antibodies identified nuclear proteins of the correct size in BeWo and HEK-293 nuclear extracts (*Fig. 2.4B*). We further tested the specificity of these antibodies by examining protein extracted from HEK-293 cells after siRNA knockdown; densitometry revealed a reduction of ZSCAN5A, 5B, and 5D proteins by 84%, 67%, and 80%, respectively (*Fig. 2.4B; Table A.3*), confirming antibody specificity as well as providing additional support for the functional efficiency of siRNA knockdowns.

We used the antibodies to human ZSCAN5A, ZSCAN5B, and ZSCAN5D for ChIP-seq in chromatin prepared from HEK-293 and BeWo cells, with chromatin from BeWo cells yielding by far the best results. ZSCAN5B peaks in BeWo chromatin were particularly clear, with very little background and strong enrichment in a limited number of genomic positions (a total of 672 peaks; 225 of which were detected with MACS software at a false discovery rate (fdr)=0; *Table A.2*). ZSCAN5A and ZSCAN5D ChIP experiments were plagued by a higher rate of background but also included larger numbers of highly and clearly enriched peaks. ZSCAN5A and ZSCAN5D ChIP experiments in chromatin from the HEK-293 cell line were not successful, but ZSCAN5B ChIP yielded a small numbers of clear peaks in this cell line. Because antibodies for all three proteins gave excellent results in BeWo chromatin, we focused on results from BeWo ChIP datasets for most types of peak analysis and used the ZSCAN5B HEK-293 ChIP experiments primarily for cell-to-cell comparisons and for functional studies including association of peaks with siRNA knockdown DEGs.

ZSCAN5B preferentially binds tDNA in human and mouse cells

We noticed immediately that high-scoring ZSCAN5B peaks in both BeWo and HEK-293 chromatin coincided significantly with tRNA genes. Statistical analysis confirmed that the enrichment for tDNA sequences in the ZSCAN5B datasets, especially the fdr=0 subset of peaks, was very high (p=0; *Table 2.2*). Of all 672 ZSCAN5B BeWo peaks, 240 overlapped with tDNA sequences (*Table A.2*). In addition to these shared peaks, 15 HEK-293 peaks occupied tDNAs located directly adjacent to one bound by ZSCAN5B in BeWo chromatin in the same cluster, and 3 additional isolated tDNAs were detected uniquely in HEK-293 ChIP in the MACs analysis. The ZSCAN5B-bound tDNAs in both cell types correspond to a variety of different amino acids and codons without obvious enrichment of a particular type (*Table A.2, Table A.4*).

To complement the human ZSCAN5B BeWo dataset, we also used the mouse Zscan5b antibody for ChIP-seq in chromatin isolated from dissected mouse fetal placenta. We identified 118 peaks, including many overlapping the homologs of the same tDNA sequences as those detected in ChIP experiments with human ZSCAN5B in syntenically homologous positions (*Table A.5*). These data confirmed that ZSCAN5B favors a specific subset of the same conserved tDNAs in different human cell types and in mouse placental tissue.

Almost all of the highest-scoring ZSCAN5B and mouse Zscan5b peaks overlapped tDNAs, but ChIP-seq also detected enrichment in other types of Pol III transcripts including Vault RNA, 7SLRNA, and U6 snRNA; 200 of the 225 BeWo peaks detected with MACs at fdr=0 overlapped with this larger collection of Pol III genes. Accordingly, ZSCAN5B peaks were enriched for many of the same classes of RNA genes found to be occupied by Pol III (measured by ChIP with an antibody to the RPC155 subunit) and TFIIIC as reported by Moqtaderi and colleagues (Moqtaderi et al., 2010) (*Table 2.2*). The small number of peaks that did not overlap with RNA Pol III genes were also interesting. For example, several individual peaks with high ChIP enrichment were found to overlie MIR and Alu SINE repeats, which are evolutionarily derived from tRNA and 7SL RNA, respectively (Ullu and Tschudi, 1984). One particular example, identified in human ZSCAN5B ChIP in BeWo and HEK-293 chromatin as well as mouse Zscan5b ChIP in fetal placenta corresponds to a MIR repeat located

approximately 1 kb downstream of the promoter of *POLR3E*, encoding the RPC5 subunit of Pol III (*Table A.2*). This particular MIR element has been shown to function as an enhancer for the *POLR3E* gene (Canella et al., 2012) suggesting that ZSCAN5B might regulate the expression of the Pol III subunit gene in some cellular contexts.

ZSCAN5A and ZSCAN5D binding sites are also enriched in Pol III transcripts

ZSCAN5A and ZSCAN5D binding sites were also highly enriched in tDNA sequences (*Table 2.2*). In fact, the three proteins appear to co-occupy many tDNA sites or to occupy neighboring tDNAs within the same genomic clusters although with different relative efficiencies. One particularly interesting set of examples is illustrated in Figure 5A; the clustered tDNAs in this region are differentially marked by the human ZSCAN5 proteins. These tDNAs, which surround the *ALOXE3, HES7* and *PER1* genes, have been shown to serve as anchors of local chromatin loops and to function as insulators in human cells (Raab *et al.*, 2012). tDNAs throughout the genome also displayed ZSCAN5 protein-specific peaks and for ZSCAN5B, cell type-specific enrichment patterns across the genome (*Table A.2*). These data suggest that the ZSCAN5A, B, and D proteins all bind to tDNA sequences but have different locus preferences within the same cell type; ZSCAN5B also clearly binds to specific tDNA loci more or less efficiently depending on the cellular context.

Motif analysis reveals binding preferences for ZSCAN5 proteins

ZSCAN5A and ZSCAN5B proteins bind G/C rich motifs

We used the MEME suite (Bailey *et al.*, 2009; Machanick and Bailey, 2011) to search for enriched sequence motifs enriched at the summits of the highest-scoring ZSCAN5A, ZSCAN5B and ZSCAN5D peak regions (see Methods). The analysis of ZSCAN5B peak summit regions revealed clear enrichment for a sequence including the TFIIIC-binding B box as the top-scoring, centrally located motif; Zscan5b peaks in mouse fetal placenta chromatin also yielded a very similar extended, central B Box motif (*Fig. 2.5B*). The A box motif typical to Pol III type 2 promoters was also identified as highly enriched in ZSCAN5B ChIP experiments although the motif was not central to the

peaks (not shown). It was not surprising to identify these motifs prominently in the ZSCAN5B peak datasets, since the A and B box elements in tDNAs are very distinct and well conserved between sites. However, the detected motifs also included less distinct G/C-rich DNA sequences extending beyond the B box sites (*Fig. 2.5B*). These extended motifs suggested that the true ZSCAN5B binding sites might be located in G/C-rich DNA surrounding the A and B box motifs.

To gain more detailed information about the binding motif, we examined 133 highly enriched (MACS enrichment factor or ef $\geq$10) ZSCAN5B BeWo peaks that were not associated with tDNAs. MEME analysis identified a long (29 bp) G-rich motif located centrally within the peaks. Adjusting MEME parameters to search for shorter motifs (since with 5 zinc fingers, ZSCAN5B is expected to bind at most to a 15 nt target region) identified two slightly different G/C-rich portions, which we call 5B_M1 and 5B_M2 (*Fig. 2.5C*). A similar G/C-rich motif identified as highly enriched at centers of ZSCAN5A peak summits, bears a striking resemblance to the known binding motif for transcription factor and chromatin organizer, CTCF (*Fig. 2.5D*). The B Box-containing motif was also detected as enriched in ZSCAN5A peak summits but this motif was not detected as centrally located (not shown). Thus both ZSCAN5A and ZSCAN5B ChIP experiments identified G/C-rich motifs as potential binding sites.

To ask whether the ZSCAN5B protein binds to the G-rich motifs, we selected one ZSCAN5B peak region for the "supershift" version of electrophoretic mobility shift assays (EMSA). The tDNAs are repetitive and mostly occupied by the very large Pol III protein complex; both properties complicate EMSA experiments. Therefore, we focused particularly on a non-tDNA peak uniquely detected with high efficiency by ZSCAN5B, which is located within an intron of the *STAP2* gene (chr19:4328490-4328689). The peak summit region contains side-by-side high-scoring matches to 5B_M1 and 5B_M2 motifs. Biotin-labeled oligonucleotides were designed to span segments of the two peak regions including the motifs for EMSA testing (*Fig. 2.6A*).

Both double-stranded oligonucleotides from the *STAP2* intronic region were shifted after addition of the protein extract, although the 5B_M2 oligonucleotide and a longer oligonucleotide containing both 5B_M1 and 5B_M2 were shifted much more intensely. Addition of the ZSCAN5B antibody caused the smallest of three "shift"

complexes for 5B_M1 and the longer oligonucleotide with both motifs (arrowheads in *Fig. 2.6B*) to be supershifted efficiently. These data indicated that the preferred binding sequence for ZSCAN5B corresponds to the G/C-rich sequence commonly identified in the non-tRNA binding peaks.

ZSCAN5D binding is enriched at MIR SINEs and LINE2-associated ETC motifs

ZSCAN5D peaks did overlap with ZSCAN5A and ZSCAN5B at tDNAs, but the highest-scoring peaks for the ZSCAN5D antibody were distinctly bound. For example, more than half (56%) of the ZSCAN5D fdr=0 peaks mapped within 5 kb of a transcription start site (TSS); a similar count yielded only approximately 20% of comparable ZSCAN5A and ZSCAN5B peaks. Enriched repetitive element classes also distinguished ZSCAN5D peaks from those associated with ZSCAN5A or ZSCAN5B; instead ZSCAN5D more closely resembled TFIIIC ChIP binding peaks and ETC site enrichments (*Table 2.2*). Furthermore, analysis of ZSCAN5D binding peaks revealed a distinct set of enriched motifs. The first and most highly enriched motif identified in this peak dataset (5D_M1) corresponded to a novel sequence with a short GA-rich segment (*Fig. 2.5E*). The second significantly enriched motif found at the center of predicted ZSCAN5D summits, or 5D_M2 is notable because it is virtually identical to a novel ETC-associated motif identified previously in human cells (Moqtaderi et al., 2010) (*Fig. 2.5E*). The two ZSCAN5D peak-enriched motifs were mostly distributed in distinct peak summits, although several ChIP summits contained both predicted motifs in close proximity.

Closer inspection revealed that 5D_M1 mapped frequently within annotated MIR SINE repeats, while 5D_ M2 was contained within LINE2 (L2) elements; both of these repeat types were highly enriched in the ZSCAN5D ChIP peak dataset as well as ETC sites (*Table 2.2*). It is interesting to note that this L2-associated ETC motif is G/C rich and does bear some similarity to the G/C-rich motifs detected with ZSCAN5B and ZSCAN5A (*Fig 2.5 C-E*). These data suggest that despite fingerprint divergence, the primate-specific ZSCAN5 proteins likely recognize G/C-rich sites that are similar to the sites preferentially bound by parental ZSCAN5B.

ZSCAN5 protein binding influences expression of Pol III and nearby Pol II genes

To ask whether ZSCAN5 protein binding might affect transcription of bound Pol III or nearby Pol II genes, we examined ZSCAN5 ChIP peaks that either mapped within or flanked DEGs detected after siRNA knockdown. Since knockdown data could only be reliably supplied by experiments in HEK-293 cells and ChIP was only successful in HEK-293 chromatin with the ZSCAN5B antibody, we focused primarily on peaks and nearby genes associated with ZSCAN5B in that cell line.

ChIP data revealed that Pol III gene *RMRP* was bound internally by ZSCAN5B in both HEK-293 and BeWo cells and also at lower efficiency by ZSCAN5A in BeWo chromatin (*Table A.2*). This gene, encoding the RNA component of mitochondrial RNase P, represents one of the very few unique-sequence Pol III genes, a fact that permitted accurate measurement of *RMRP* transcript levels in RNA-seq experiments. *RMRP* was significantly up-regulated in both *ZSCAN5B* and *ZSCAN5A* siRNA experiments suggesting that ZSCAN5 protein binding suppresses expression of the gene (*Fig. 2.7*). Most other Pol III genes are highly repetitive and the RNA products fold into very stable secondary structures, so that expression is therefore difficult to measure accurately. However, primer sets that can detect a small number of tRNAs uniquely have been reported (Canella et al., 2012). Of the genes expressing these tRNAs, one was occupied by ZSCAN5A and ZSCAN5B in BeWo cells and ZSCAN5B in HEK-293; a second gene, encoding to a tRNA-Arg and located on chr9, was occupied by ZSCAN5A and ZSCAN5B in BeWo cells but not detected as bound by ZSCAN5B in HEK-293 cells (*Table A.2*). The other tRNA genes for which unique primer sets have been reported were not occupied by ZSCAN5A or ZSCAN5B proteins in ChIP experiments in either cell type. We selected two tRNA primer sets of each type (bound and unbound) to test expression after ZSCAN5 gene knockdown with qRT-PCR in HEK-293 cells, along with primers for *RMRP*. *RMRP* and the tRNA expressed from the tRNA bound in both cell types were up-regulated after siRNA knockdown; the tRNA expressed from the chr9 gene not occupied by ZSCAN5B in HEK-293 cells was up-regulated only after knockdown of ZSCAN5A. In contrast, the expression of unbound tDNAs remain unchanged in expression (*Fig 2.7*). Although further studies will be required to test the

broader generality of this trend, these data support the hypothesis that ZSCAN5 proteins negatively regulate transcription of Pol III genes to which they are bound.

Next we examined expression status of Pol II genes with transcription start sites (TSS) located very near (within 2 kb of) ZSCAN5B HEK-293 peaks. We identified 13 *ZSCAN5B* DEGs of this type, including seven genes –*HES7, TRIM7, NDUFS7, TIA1, C16orf13,* and *C17orf59,* and *Slc27a4* - located adjacent to tDNA peaks. Of these seven genes, *TIA1* was down-regulated after *ZSCAN5B* knockdown while all others were up-regulated. In the larger set of 13 adjacent genes including those adjacent to unique sequence peaks, all but *TIA1* and *PPP1R12A* were up-regulated in the knockdown experiment suggesting that ZSCAN5B binding represses most adjacent genes.

However, looking at all DEGs with promoters located further from ZSCAN5B HEK-293 peaks (within 20 kb, including both tDNA and non-tDNA peaks) this picture was more mixed, with 25 genes up-regulated and 13 genes down-regulated after ZSCAN5B siRNA knockdown (*Table A.2*). These data suggest a general trend toward repression, but also indicate that the effects of ZSCAN5 protein binding on nearby genes may not be as simple as a direct transcriptional repression or activation and may involve more complicated mechanisms.

DISCUSSION

Experiments reported here show that the ZSCAN5 family genes are actively expressed in rapidly dividing cell populations and that in those cells, the ZSCAN5 proteins bind preferentially to Pol III transcription units, especially tDNAs, and related repetitive elements. The conserved progenitor of the primate family, ZSCAN5B, binds with high specificity to a particular subset of tDNA sequences in two human cell lines and strikingly, to the same subset of homologous tDNAs in chromatin of mouse fetal placenta. Although it is expressed at higher levels in human cells and displays a wider range of binding sequences, primate-specific ZSCAN5A also binds preferentially to tDNA, co-occupying many tDNA sites with ZSCAN5B. A functional cooperation between ZSCAN5A and ZSCAN5B proteins was further suggested by the high degree of similarity, in terms of gene identities as well as their up- or down-regulation, between the

DEGs that were detected after *ZSCAN5A* and *ZSCAN5B* siRNA knockdown. Since the human genes encode nearly identical SCAN elements, we surmise that it is likely that ZSCAN5 proteins interact as heterodimers through those domains.

We identified a number of DEGs, including a mixture of up- and down-regulated genes, that flanked or contained ChIP peaks for ZSCAN5A, ZSCAN5B, or both; of the Pol II-expressed DEGs located within 2 kb of a ZSCAN5B HEK-293 ChIP peak, all but two were significantly up-regulated after *ZSCAN5B* siRNA knockdown suggesting that ZSCAN5B binding acts to repress transcription of those genes. The mechanism through which ZSCAN5 protein binding could regulate those genes will require further, in-depth study. However, based on our data and available data from other groups, we hypothesize that the effects on Pol II gene transcription may be a secondary effect of regulating the Pol III genes.

Many previous studies have pointed clearly to a relationship between Pol III transcriptional activity and the expression of nearby Pol II genes. For example, in budding yeast, actively expressed tDNAs have long been known to exert repressive effects on nearby Pol II promoters, a phenomenon referred to as “tRNA gene-mediated silencing” (Hull et al., 1994; Kinsey and Sandmeyer, 1991). However available data from other species suggest a more complex relationship. In particular, tDNAs, MIR repeats, and other Pol III-transcribed sequences can serve enhancer-blocking or barrier-insulating functions, with varying effects on flanking Pol II genes. For example, higher levels of Pol III recruitment to promoter-linked MIR elements resulted in significant enhancement of *CDKN1A* and *GDF15* gene expression in human cells. In the case of *GDF15*, Pol III activity was shown to be required for formation of a chromatin loop that activates transcription of the Pol II gene (Lee et al., 2015). However, loop structures can also lead to gene repression through enhancer blocking and other mechanisms (Comet et al., 2011; Jiang and Peterlin, 2008; Kirkland et al., 2013; Li et al., 2008; Van Bortle and Corces, 2012). Our data suggest that ZSCAN5B binding represses Pol III transcription, and we speculate that through this mechanism, ZSCAN5B modulates the formation of chromatin loops at bound tDNAs and interacting sites. If this hypothesis is correct, the positive or negative effects of ZSCAN5 binding on nearby Pol II genes could be subtle, depending

on distance from the peak and on the particular types of interactions between enclosed genes and the anchors of the modulated chromatin loops.

Whether directly or indirectly, *ZSCAN5A* and *ZSCAN5B* knockdown robustly affected the expression of large sets of functionally coherent genes, which together provide clues to the ZSCAN5 proteins' ultimate biological functions. ZSCAN5A has been previously reported as one of 850 genes differentially regulated during the cell cycle, with peak expression during the G2/M transition (Whitfield et al., 2002), and consistent with this finding, knockdown of *ZSCAN5A* predicted that the protein negatively regulates processes involved in chromosome segregation and metaphase-anaphase transition. Positively regulated genes, by contrast, were enriched in functions related to cellular differentiation, with genes active in development of lung, gut, blood vessels, neurons, immune cells, and several other tissues in which the ancestral mouse gene is actively expressed during embryogenesis. Together these data suggest that *in vivo,* ZSCAN5A may regulate cell cycle exit and differentiation in a variety of precursor cell types.

ZSCAN5B knockdown revealed additional cellular functions, related to tRNA maturation and modification events. These predicted functions fit well with the hypothesis that ZSCAN5B binding also negatively regulates the synthesis of tRNA genes. Here we should note that in addition to the essential functions tRNAs serve in protein synthesis, tRNA processing generates fragments that functions as microRNAs and serve independent signaling functions; the regulation of tRNA processing, *per se,* may thus exert profound influences on gene expression and cellular state (Kirchner and Ignatova, 2015).

Indeed, many of the biological functions suggested for ZSCAN5 proteins may be encapsulated by a brief summary of known functions for *RMRP,* a Pol III-transcribed gene and directly repressed target of ZSCAN5A and ZSCAN5B. RMRP plays a critical role in processing of ribosomal RNAs, mitochondrial tRNAs, and microRNA precursors (Martin and Li, 2007) and also processes cellular mRNAs, most notably degrading Cyclin B2 mRNA to permit cell cycle progression at the end of mitosis (Gill et al., 2004). *RMRP* mutations are associated with cartilage hair hypoplasia-Anauxetic dysplasia (CHH-AD) spectrum disorders, associated with a range of symptoms including short-limbed

dwarfism, skeletal dysplasia, hair abnormalities, immunodeficiency and bone marrow failure, gastrointestinal disorders, cognitive defects, and cancer susceptibility (Martin and Li, 2007; Thiel et al., 2007). Rather than being caused simply by loss of RNAse P function *per se*, many CHH-AD symptoms are thought to reflect malfunction of RMRP-derived microRNAs that serve to target and down-regulate essential developmental genes (Rogler et al., 2014). Disruptions in expression of this target locus could thus set off a cascade of events with significant downstream consequences.

Because we could not knock down *ZSCAN5D* with siRNA specifically, the biological functions of this primate-specific paralog remain something of a mystery. However, our data show that ZSCAN5D protein displays highest preference for tDNA-derived MIR SINE sites and a subset of LINE2 elements that harbor a previously identified ETC motif (Moqtaderi et al., 2010). L2 elements have carried the transposition-deficient MIR SINEs as "hitchhikers", and remnants of the two repeats can often be found closely juxtaposed (Gilbert and Labuda, 2000). The high enrichment of both ZSCAN5D and the human ETCs in MIR SINE and L2 repetitive elements (*Table 2.2*), suggests that a subset of ETC sites were distributed as MIR-associated L2 repeats during early mammalian evolution, and that these ETCs have evolved as preferred binding sites for ZSCAN5D. Further studies will be required to address these hypotheses definitively.

In closing, we note that this TF gene family was elaborated, first in eutherians and further in early primate lineages, to regulate one of the most ancient mechanisms known to control three-dimensional chromatin architecture in eukaryotic cells. We surmise that *ZSCAN5B* evolved in early eutherian history to add a novel layer of regulation on a subset of Pol III-transcribed genes, modulating their transcription and their chromatin-organizing functions. The new primate paralogs diverged in fingerprint structure but retained a preference for a similar G-rich binding motif to co-bind with ZSCAN5B at many tDNA sites, likely dimerizing through their nearly identical SCAN domains. Through modulation of Pol III transcription and position effects on neighboring genes, our data suggest that human ZSCAN5A and ZSCAN5B collaborate in control of noncoding RNA processing, cell cycle progression and differentiation in many tissues. Although functional understanding of ZSCAN5D is complicated and still incomplete, the observation that this primate-specific protein binds to MIR-repeat derived ETCs offers a

potentially valuable clue to the evolution of mammalian chromatin structure and deserves further investigation.

MATERIALS AND METHODS

ETHICS STATEMENT

This investigation has been conducted in accordance with the ethical standards and according to the Declaration of Helsinki and according to national and international guidelines. All animal work was reviewed and approved by the University of Illinois IACUC committee under protocol number 15425.

RNA PREPARATION AND QUANTITATIVE RT-PCR

Total RNA was isolated from cell lines and tissues using TRIzol (Invitrogen) followed by 30 minutes of RNase-free DNaseI treatment (NEB) at 37ºC and RNA Clean & Concentrator™-5 (Zymo Research). 2 µg of total RNA was used to generate cDNA using Superscript III Reverse Transcriptase (Invitrogen) with random hexamers (Invitrogen) according to manufacturer's instructions.

Resulting cDNAs were analyzed of transcript-specific expression through quantitative reverse-transcript PCR (qRT-PCR) using Power SYBR Green PCR master mix (Applied Biosystems) with custom-designed primer sets (*Table A.6*) purchased from Integrated DNA Technology. Relative expression was determined by normalizing the expression of all genes of interest to either human or mouse Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, zeta polypeptide (*YWHAZ*) expression (∆Ct) as described (Eisenberg and Levanon, 2003).

IN SITU HYBRIDIZATION

Mouse embryos were collected, paraffin embedded, sectioned and hybridized with an antisense RNA probe, essentially as previously described (Kim et al., 2001). We generated *In Situ* hybridization (ISH) probes correspond to nucleotides of the mouse *Zscan5b* cDNA sequence (NM_133204), and to human ZSCAN5B (NM_001080456) (see *Table A.6* for probe sequences). Probes were cloned into the pGEM-T vector

(Promega) and sequence-validated before being used for ISH. The reverse primer included a T7 promoter sequence to permit antisense RNA generation using the Roche DIG RNA Labeling kit (SP6/T7) (Roche Applied Science) according to manufacturer's instructions. To prepare human tissue arrays, paraffin blocks containing formalin-fixed tissues from normal anonymous adult donors were purchased from NoblePath Inc. Tissue microarray (TMA) blocks were generated using a Tissue Arrayer (Beecher Instruments). The pre-cut paraffin tissue sections were checked by H&E staining for tissue index selection. Forty-one 1.5mm diameter cores were included in the arrays, with 2 cores included to represent different regions of some tissues (x2). High quality 4-micron sections were generated using a Leica ST 2155 microtome. Slides were baked at 41º C overnight and stored at -20ºC until use. Slides were pretreated and hybridizations were performed as described previously (Kim *et al.*, 2000). Sections were mounted using Vectashield Mounting Medium with DAPI. Fluorescent images were reviewed using an Olympus BX60 microscope and captured by an Olympus CC-12 digital camera.

CELL CULTURE AND TRANSFECTIONS

HEK-293 (ATCC, CRL-1573), Neuro-2a (ATCC, CCL-131), and BeWo (ATCC, CCL-98) cell lines were obtained from the American Type Culture Collection. HEK-293 and Neuro-2a cells were maintained in Dulbecco's Modified Eagle's Medium (DMEM) containing 2 mM L-glutamine, 10% fetal bovine serum (FBS), 1X Pen Strep, and BeWo cells in DMEM/F12K containing 2 mM L-glutamine, 10% FBS, 1X NEAA, 1X Pen Strep, incubated at 37 °C in 5% CO2.

For siRNA knockdown, approximately $4.5x10^5$ HEK-293 cells were seeded to 6-well plates 24 hours before transfection. Cells were treated with 10 nM of siRNA specific to each ZSCAN5A (si4: SI00779436, si5: SI04221826, Qiagen) or ZSCAN5B (si1: SI00503300, Qiagen), or ZSCAN5D (si2: SI02804774, Qiagen) with a scrambled negative control (Silencer negative control No.1 siRNA, Ambion) for 48 hours using Lipofectamine RNAi MAX transfection reagent (Invitrogen) according to manufacturer's instructions. (See *Table A.6*)

RNA-SEQ AND COMPUTATIONAL ANALYSIS

48 hours after siRNA treatment, total RNA was prepared and tested for quality using an Agilent BioAnalyzer and Illumina libraries generated using the KAPA Stranded mRNA-Seq kit with mRNA Capture Beads (Kapa Biosystems, KK8420). Sequencing was performed on an Illumina Hi-Seq 2000 instrument at the University of Illinois Roy J. Carver Biotechnology sequencing facility, to yield 60-65 million reads per sample. The data have been submitted to the Gene Expression Omnibus database (accession numbers, in progress).

RNA-seq data were analyzed using the Tophat-Cufflinks Suite of tools (Trapnell et al., 2012). For *ZSCAN5A* knockdown, expression results from si4 and si5 were analyzed as a group in comparison with the scrambled control. Genes identified as differentially expressed with $p \leq 0.05$ (after Benjamini-Hochberg correction for multiple testing) compared to the negative control-treated samples were considered for further analysis. For *ZSCAN5B* knockdown, which was effective only for a single siRNA design, we considered all genes with expression levels of at least 1 FPKM in at least one sample and considered genes with $\geq$ 1.5 X fold change relative to scrambled control as DEGs. siRNA up-regulated and down-regulated genes were analyzed for function separately using the DAVID (Huang et al., 2007a) functional clustering algorithm with default settings.

PROTEIN PREPARATION, WESTERN BLOTS, AND ANTIBODIES

Nuclear Extracts were prepared with NucBuster^{TM} Protein Extraction Kit (Novagen) and measured by Bradford-based assay (BioRad). The extracts were stored at -80^{o}C and thawed on ice with the addition of protease inhibitor Cocktail (Roche) directly before use. 15 µg of nuclear extracts were run on 10% acrylamide gels and transferred to hydrophobic polyvinylidene difluoride (PVDF) membrane (GE-Amersham, 0.45 µm) using BioRad Semi-dry system, then visualized by exposure to MyECL Imager (Thermo Scientific).

Rabbit polyclonal antibodies were generated by injection of synthetic peptides corresponding to the tether regions of ZSCAN5A (DLVRAKEGKDPPKIAS) and mouse Zscan5b proteins (CPEPANPQPEKQVDSL); peptide synthesis, antibody production,

and affinity purification of antibodies against the purified peptide epitope we carried out by Abgent Inc. ZSCAN5B (sc-249845, Santa Cruz Biotechnology) and ZSCAN5D (ARP47809_P050, Aviva Systems Biology) antibodies were obtained from commercial sources. Antibody preparations were tested for protein specificity and efficiency by Western blot staining along with anti-TATA binding protein control antibody (1TBP18, Abcam)

CHROMATIN IMMUNOPRECIPITATION

Chromatin immunoprecipitation was carried out as essentially as described (Kim et al., 2003) with modifications for ChIP-seq. Chromatin was prepared from HEK-293, BeWo, and Neuro-2a cell lines. About 1.0 x 10^6 Cells were fixed in PBS with 1% formaldehyde for 10 minutes. Fixing reaction was stopped with addition of Glycine to 0.125M. Fixed cells were washed 3x with PBS+Protease inhibitor cocktail (PIC, Roche) to remove formaldehyde. Washed cells were lysed to nuclei with lysis solution – 50 mM Tris-HCl (pH 8.0), 2 mM EDTA, 0.1% v/v NP-40, 10% v/v glycerol, and PIC – for 30 minutes on ice. Cell debris was washed away with PBS with PIC. Nuclei were pelleted and flash-frozen on dry ice.

To obtain chromatin samples from mouse fetal placenta, pregnancies were identified in C57BL/6J females by detection vaginal plugs (day of plug detection is E0.5), and placentas were dissected at E17.5. After removing maternal decidua, samples were homogenized in cold PBS containing PIC using a loose-fitting Dounce homogenizer. Debris was removed using cell strainer (Fisher Scientific), and single cell suspension was washed twice with cold PBS. Viable cells were counted using hemocytometer and crosslinked with 4% paraformaldehyde (Electron Microscopy Sciences) for 10 minutes at room temperature before nuclear isolation.

Cross-linked chromatin was prepared and sonicated using Bioruptor UCD-200 in ice water bath to generate DNA fragments 200-300 bp in size. Twenty micrograms of each antibody preparation, or 20 µg IgG for mock pulldown controls, were incubated with chromatin prepared from nuclei of approximately 5 million cells.

DNA was released and quantitated using Qubit 2.0 (Life Technologies) with dsDNA HS Assay kit (Life Technologies, Q32854), and 15 ng of DNA was used to

generate libraries for Illumina sequencing. ChIP-seq libraries were generated using KAPA LTP Library Preparation Kits (Kapa Biosystems, KK8232) to yield two independent ChIP replicates for each antibody. We also generated libraries from sonicated genomic input DNA from the same chromatin preparations as controls. Libraries were bar-coded with Bioo Scientific index adapters and sequenced to generate 15-23 million reads per duplicate sample using the Illumina Hi-Seq 2000 instrument at the University of Illinois W.M. Keck Center for Comparative and Functional Genomics according to manufacturer's instructions. Separate ChIP preparations were generated for qRT-PCR validation experiments; in this case, released DNA was amplified by GenomePlex® Complete Whole Genome Amplification (WGA) Kit (Sigma, WGA2).

CHIP-SEQ DATA ANALYSIS

Human ZSCAN5A, ZSCAN5B, and ZSCAN5D ChIP-enriched sequences as well as reads from the input genomic DNA were mapped to the HG19 human genome build, and Mouse Zscan5b ChIP reads and input mapped to the mouse Mm9 genome assembly, using Bowtie 2 software (Langmead et al., 2009) allowing 1 mismatch per read but otherwise using default settings. Bowtie files were used to identify peaks in human ChIP samples using MACS software (version 14.2) (Zhang et al., 2008), with default settings. All samples, including the Mouse Zscan5b and human ZSCAN5B ChIP samples isolated from HEK-293 cells, both of which had higher levels of background compared to other samples generated, were also analyzed using the more sensitive HOMER software package (http://homer.salk.edu/homer/ngs/index.html) using default conditions for the TF setting and false discovery rate cutoffs of 0.1 (mouse Zscan5b ChIP) or 0.01 (for ChIP in human HEK-293 and BeWo chromatin). After comparison of the individual files, sequence reads from the two separate ChIP libraries were pooled and a final peak set determined in comparison to genomic-input controls. Peaks were mapped relative to nearest transcription start sites using the GREAT program (McLean et al., 2010). Peak locations provided by Moqtaderi *et al.* (Moqtaderi et al., 2010) and Oler *et al.* (Oler et al., 2010) were converted to Hg19 coordinates using the Liftover utility provided by the UCSC genome browser (https://genome.ucsc.edu/), and overlaps between these peaks and ZSCAN5 ChIP data were determined. DEGs and ChIP peaks, with overlaps to these

other datasets and repeat-sequence mapping are summarized in *Table A.2* for human cell types and *Table A.5* for mouse fetal placenta ChIP.

REPETITIVE ELEMENT OVERLAP ANALYSIS

To identify enrichment or under-representation of repetitive element types or families in the ChIP-peak datasets, we used a method modified from that described by Cuddapah and colleagues (Cuddapah et al., 2009). Human repeat data were retrieved on 11/25/2013 as the RepeatMasker Table (www.repeatmasker.org) from USCS's table browser (genome.ucsc.edu) (Karolchik et al., 2004) with the following parameters: assembly= 'Feb. 2009 (GRCh37/hg19)', group= 'Variation and Repeats', track= 'RepeatMasker', table= 'rmsk', region= 'genome', output format= 'BED – Browser extensible data'. The human chromosome sizes required for the analysis were retrieved on 2/20/2014 from the hg19.chromInfo table of the UCSC public database (Kuhn et al., 2013). We examined overlap between genome coordinates of repeat element features and 100 bp intervals surrounding the summits of peaks determined by MACS software from ZSCAN5A, ZSCAN5B, and ZSCAN5D ChIP experiments (986 peaks with FDR=0 or effective fold change ≥ 15 for ZSCAN5A_BeWo_ef15; 225 peaks with FDR=0 for ZSCAN5B_BeWo; 1885 peaks with FDR=0 for ZSCAN5D_BeWo, and all 102 peaks reported by MACS for ZSCAN5B_HEK-293) using the BEDTools intersect function (Quinlan and Hall, 2010). Overlaps were also determined for peak sets from Moqtaderi and colleagues (Moqtaderi et al., 2010) after applying the UCSC liftover ultility to identify coordinates in the human hg19 genome build (ETC with 1865 200 bp peaks; RPC155 with 1518 200 bp peaks; and TFIIIC with 5472 200 bp peaks).

For each peak set 500 random sets of the same number and peak size were generated by the BEDTools random function, and overlaps between these random peak sets and repeats were counted for each of the 500 random sets. For each repeat element and family the average overlap count of the random sets and the standard deviation was determined. Then for each repeat element and family a Z-score was calculated using the overlap count of the peak set, and the average overlap count and standard deviation of the random sets.

If the overlap count of the peak set was less than or equal to the average of the random sets z was calculated as: $z = \frac{(overlap\ count\ of\ peak\ set) - (average\ overlap\ count\ of\ random\ sets)}{standard\ deviation\ of\ the\ overlap\ count\ of\ random\ sets}$. If the overlap count of the peak set was greater: $z = \frac{(average\ overlap\ count\ of\ random\ sets) - (overlap\ count\ of\ peak\ set)}{standard\ deviation\ of\ the\ overlap\ count\ of\ random\ sets}$. The R function pnorm(z) was used to calculate a p-value to indicate if the overlap count was significantly under-represented or enriched in a ChIP-peak set when compared to the overlap counts of the random sets. Repeat families or specific elements that were significantly enriched in at least one of the ChIP peak sets are reported in *Table 2.2* and *Table A.4*, respectively, along with p-values determined for enrichments or under-representation of that family or element type in each peak set.

MOTIF ANALYSIS

To identify enriched motifs, we used sequence from a 200 bp region surrounding the predicted summits of selected peaks for analysis with MEME-ChIP with default parameters (Machanick and Bailey, 2011). Motifs displayed in Fig. 5 were identified from peaks with the following cutoffs: (panel B) All HOMER-derived peaks for mouse Zscan5b ChIP in fetal placenta chromatin; (2) All HOMER peaks with enrichment (ef) $\geq$ 20 in human ZSCAN5B ChIP with HEK-293 chromatin; and All MACS ef $\geq$ 20, fdr=0 peaks from ZSCAN5B ChIP in BeWo chromatin; (panel C) all peaks *except* those overlying tDNAs, identified with MACS in BeWo chromatin with ef >10; (panel D and E) ZSCAN5A and ZSCAN5D peaks identified with MACS at ef >20. Only the top peaks identified in each analysis are shown, as described in the text.

SUPERSHIFT ELECTROPHORETIC MOBILITY SHIFT ASSAY

Nuclear extracts from subconfluent HEK-293 cells were prepared with NucBuster™ Protein Extraction Kit (Novagen) and measured by Bradford-based assay (BioRad). Probes were synthesized as double stranded oligonucleotides by annealing 2 µg of biotin 5'-end labeled single stranded oligonucleotides and unlabeled complementary single stranded oligonucleotides (Integrated DNA Technology) in annealing buffer (10 mM Tris-HCl, pH7.5, 50 mM NaCl, 1 mM EDTA). Mixed

oligonucleotides were heated in 95ºC hot block for 10 minutes and slowly cooled down to room temperature. The sequences of the probes used were as follows: STAP2_M2 forward (5'biotin- CGGGTCGGACTCCGCCCCTGCTTCTGA-3') and reverse (5'-TCAGAAGCAGGGGCGGAGTCCGACCCG-3'); STAP2_M1 forward (5'biotin-CTGACCACGCCCCCGCGCCCACCCTCTT-3') and reverse (5'-AAGAGGGTGGGCGCGGGGGCGTGGTCAG-3'); STAP2_M2+M1 forward (5'biotin-CGGGTCGGACTCCGCCCCTGCTTCTGACCACGCCCCCGCGCCCACCCTCTT-3') and reverse (5'-AAGAGGGTGGGCGCGGGGGCGTGGTCAGAAGCAGGGGCGGAGTCCGACCCG-3'). EMSA binding reactions were performed at room temperature for 30 minutes and consisted of 3 µg of nuclear extract in 1X binding buffer, 2.5% glycerol, 5 mM $MgCl_2$, 50 ng/µl poly(dI-dC), 0.05% NP-40, and 20 ng of biotinylated DNA probes. For supershift reactions, 5 µg of antibody targeting ZSCAN5B was carefully added and mixed, and incubated for additional 30 minutes at room temperature. The mixture was run on 6% non-denaturing polyacrylamide gels in 1X Tris borate-EDTA buffer. Protein-DNA complexes were then transferred to PVDF membrane using the BioRad Semi-dry system and cross-linked using the Spectrolinker XL-1000 UC cross-linker (Spectronics Corp.). Detection of biotin-labeled DNA was performed using the LightShift chemiluminescent EMSA kit (Thermo Scientific) and visualized by exposure to MyECL Imager, a charge-coupled device camera (Thermo Scientific).

ACKNOWLEDGEMENTS

We would like to thank Dr. Elbert Branscomb and Dr. Saurabh Sinha for critical comments on the manuscript.

DISCLOSURE OF POTENTIAL CONFLICT OF INTEREST

The authors declare no conflicts of interest.

GRANT SUPPORT

This work was supported by March of Dimes Foundation, grant FY2011-393 (awarded to L.S.), and by the National Institutes of General Medical Sciences grants R01-GM078368 (L.S.) and R01-GM114341 (S.S.).

FIGURES

Figure 2.1.

Proteins	Fingerprints				
	1	2	3	4	5
Mm_Zscan5b	YSRL	QSDV	HSTG	HGNV	QGTR
Cj_ZSCAN5B	YSQL	QSDV	HSTG	HGNV	QGTR
Hs_ZSCAN5B	YSQI	QSDV	HSTG	SGNV	QGTR
Cj_ZSCAN5A	CSKI	QISF	QSYC	YGSE	RKLR
Hs_ZSCAN5A	CSKI	QISF	QSYC	YGSE	RKLR
Cj_ZSCAN5C	YGKI	QIGL	QSYC	YGSE	RATR
Hs_ZSCAN5C	YGKV	QIGF	QSYC	YANE	RATR
Cj_ZSCAN5D	YSKI	QSDV	RFSC	YKNQ	RETR
Hs_ZSCAN5D	CSKI	QSDV	RFSC	YKNE	RETY

Figure 2.1. (cont.) DNA-binding "fingerprint" patterns for primate-specific ZSCAN5 proteins

The fingerprints of the novel duplicates have diverged relative to the ancestral gene copy, but once generated, these patterns have been conserved in primate species. Amino acid residues corresponding to DNA-binding positions (1-, 2, 3, and 6 relative to the alpha helix) in each of the five zinc fingers in ZSCAN5 proteins predicted from mouse, human, and marmoset genomes are shown aligned in the 5'->3' order of the zinc fingers in each gene. Mouse and other mammals contain a single gene, *Zscan5b,* which was duplicated in early primate history to generate three new gene copies: *ZSCAN5A, ZSCAN5C,* and *ZSCAN5D* (Mm_Zscan5b: ENSMUSG00000058028, Cj_ZSCAN5B: ENSCJAG00000020279, Hs_ZSCAN5B: ENSG00000197213, Cj_ZSCAN5A: ENSCJAG00000037918, Hs_ZSCAN5A: ENSG00000131848, Cj_ZSCAN5C: ENSCJAG00000020275, Hs_ZSCAN5C: ENSG00000204532, Cj_ZSCAN5D: ENSCJAG00000014787, Hs_ZSCAN5D: ENSG00000267908).

Figure 2.2.

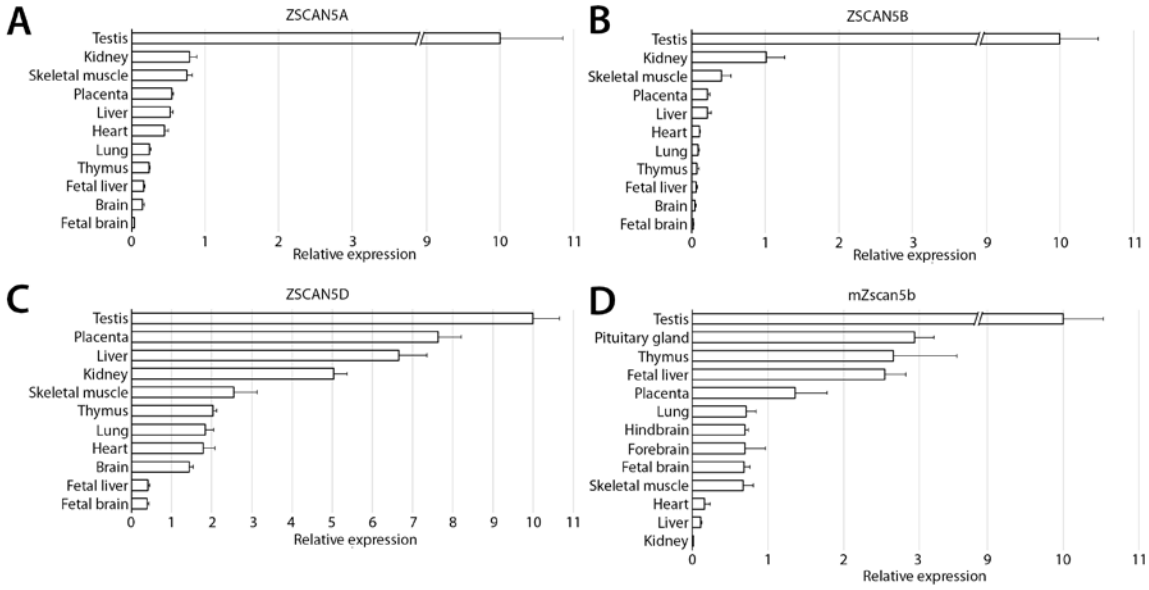

A
ZSCAN5A
Testis
Kidney
Skeletal muscle
Placenta
Liver
Heart
Lung
Thymus
Fetal liver
Brain
Fetal brain
0 1 2 3 9 10 11
Relative expression
B
ZSCAN5B
Testis
Kidney
Skeletal muscle
Placenta
Liver
Heart
Lung
Thymus
Fetal liver
Brain
Fetal brain
0 1 2 3 9 10 11
Relative expression
C
ZSCAN5D
Testis
Placenta
Liver
Kidney
Skeletal muscle
Thymus
Lung
Heart
Brain
Fetal liver
Fetal brain
0 1 2 3 4 5 6 7 8 9 10 11
Relative expression
D
mZscan5b
Testis
Pituitary gland
Thymus
Fetal liver
Placenta
Lung
Hindbrain
Forebrain
Fetal brain
Skeletal muscle
Heart
Liver
Kidney
0 1 2 3 9 10 11
Relative expression

Figure 2.2. (cont.) ZSCAN5 family exhibits overlapping but distinct patterns of tissue-specific expression in human tissues

Relative transcript levels were measured in cDNA prepared from total RNA from human tissues by qRT-PCR for (A) human *ZSCAN5A*, (B) *ZSCAN5B*, (C) *ZSCAN5D*, and from mouse tissues for (D) mouse *Zscan5b*. Relative expression was normalized for other tissues compared to expression level in testis, which was set as 10. Error bars correspond to the variance between experimental triplicates.

Figure 2.3.

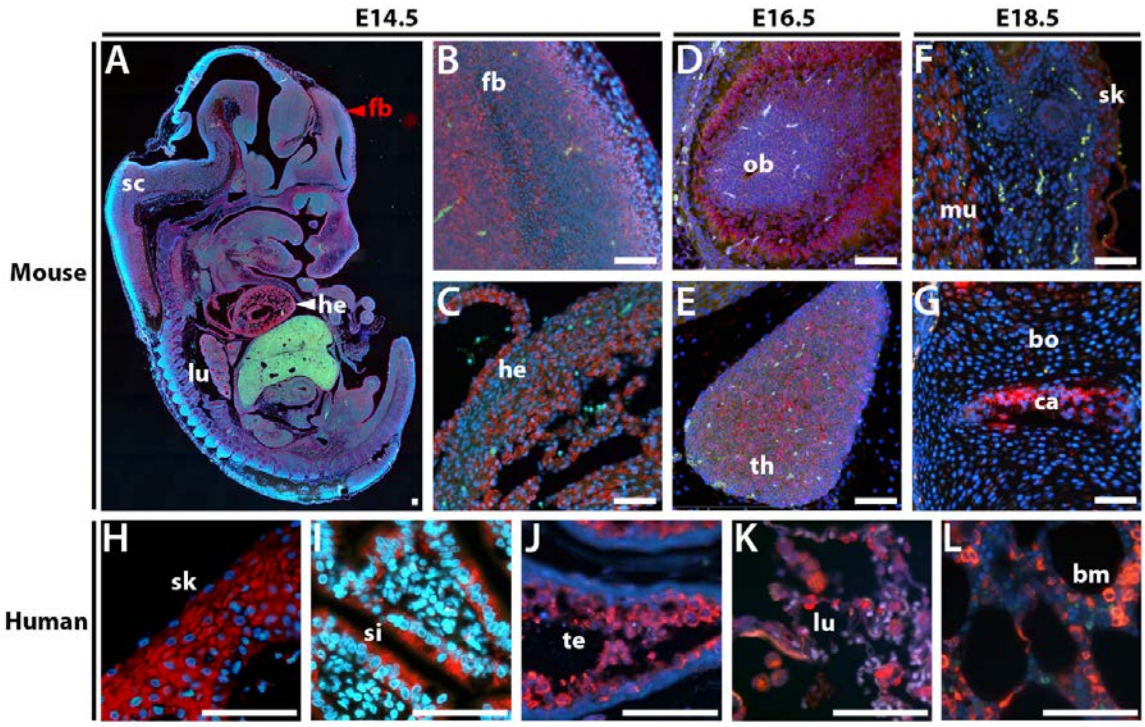

E14.5
E16.5
E18.5
Mouse
Human
A
fb
sc
he
lu
B
fb
C
he
D
ob
E
th
F
sk
mu
G
bo
ca
H
sk
I
si
J
te
K
lu
L
bm

Figure 2.3. (cont.) *In situ* hybridization in sectioned embryos and adult human tissues shows that *ZSCAN5B* expression is highest in rapidly dividing cell populations

Sagittal sections of paraffin embedded mouse embryos at E14.5, E16.5, and E18.5 were hybridized with a *Zscan5b* antisense RNA probe detected with TSA-Rhodamine (Red), against a DAPI counterstain (blue). (A) Whole sectioned E14.5 embryo shows tissues overall expression, with higher magnification panels showing expression in (B) forebrain (fb), as well as (C) heart (he) and alveoli of the lung (lu). At E16.5, high levels of expression were detected in olfactory bulb (D), and thymus (E). By E18.5, high levels of expression had faded from heart and lung but were predominant in (F) muscle (mu) and skin (sk), and (G) cartilage (ca). In human adult tissues, the *ZSCAN5B* gene was detected at particularly high levels in (H) skin epithelium (sk), (I) small intestine (si), (J) testicular spermatocytes (te) and (K) lung (lu), and bone marrow granulocytes (bm). Scale bar corresponds to 100 μm.

Figure 2.4.

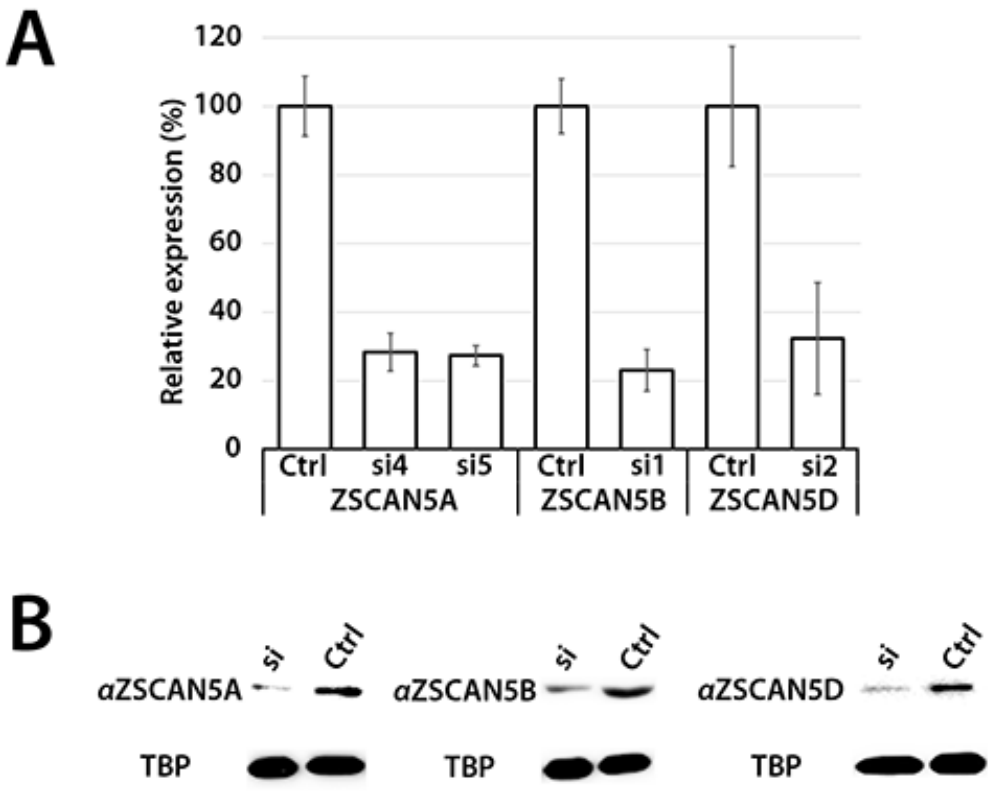

A
120
100
80
60
40
20
0
Relative expression (%)
Ctrl
si4
si5
Ctrl
si1
Ctrl
si2
ZSCAN5A
ZSCAN5B
ZSCAN5D
B
si
Ctrl
αZSCAN5A
αZSCAN5B
αZSCAN5D
TBP
TBP
TBP

Figure 2.4. (cont.) siRNA knockdown of ZSCAN5 and antibody specificity
HEK-293 cells were transfected with either 10 nM of scrambled siRNA control (Ctrl) or siRNAs targeting ZSCAN5A (si4, si5), ZSCAN5B (si1), or ZSCAN5D (si2) for 48 hours, and RNA and nuclear extracts were collected. (A) qRT-PCR of each sample showed 71.6% (si4) and 72.6% (si5) knockdown of *ZSCAN5A* transcripts, 76.9% (si1) knockdown of *ZSCAN5B* transcripts, and 67.6% (si2) knockdown of *ZSCAN5D* transcripts compared to the scrambled control. (B) To confirm knockdown of each protein and verify antibody specificities, Western blots were performed on cell extracts from cells after treatment with gene-specific siRNA or scrambled control, using antibodies targeting ZSCAN5A, ZSCAN5B, and ZSCAN5D. TATA-binding protein (TBP) was used as an internal control. Correct band sizes were detected at 56 kDa (ZSCAN5A, ZSCAN5B, ZSCAN5D), and 38 kDa (TBP). Densitometric analyses of band intensity using Image J (Schneider et al., 2012) showed 84.1% (ZSCAN5A), 67.0% (ZSCAN5B), and 80.0% (ZSCAN5D) decrease of each protein after siRNA knockdown (*Table A.3*).

Figure 2.5.

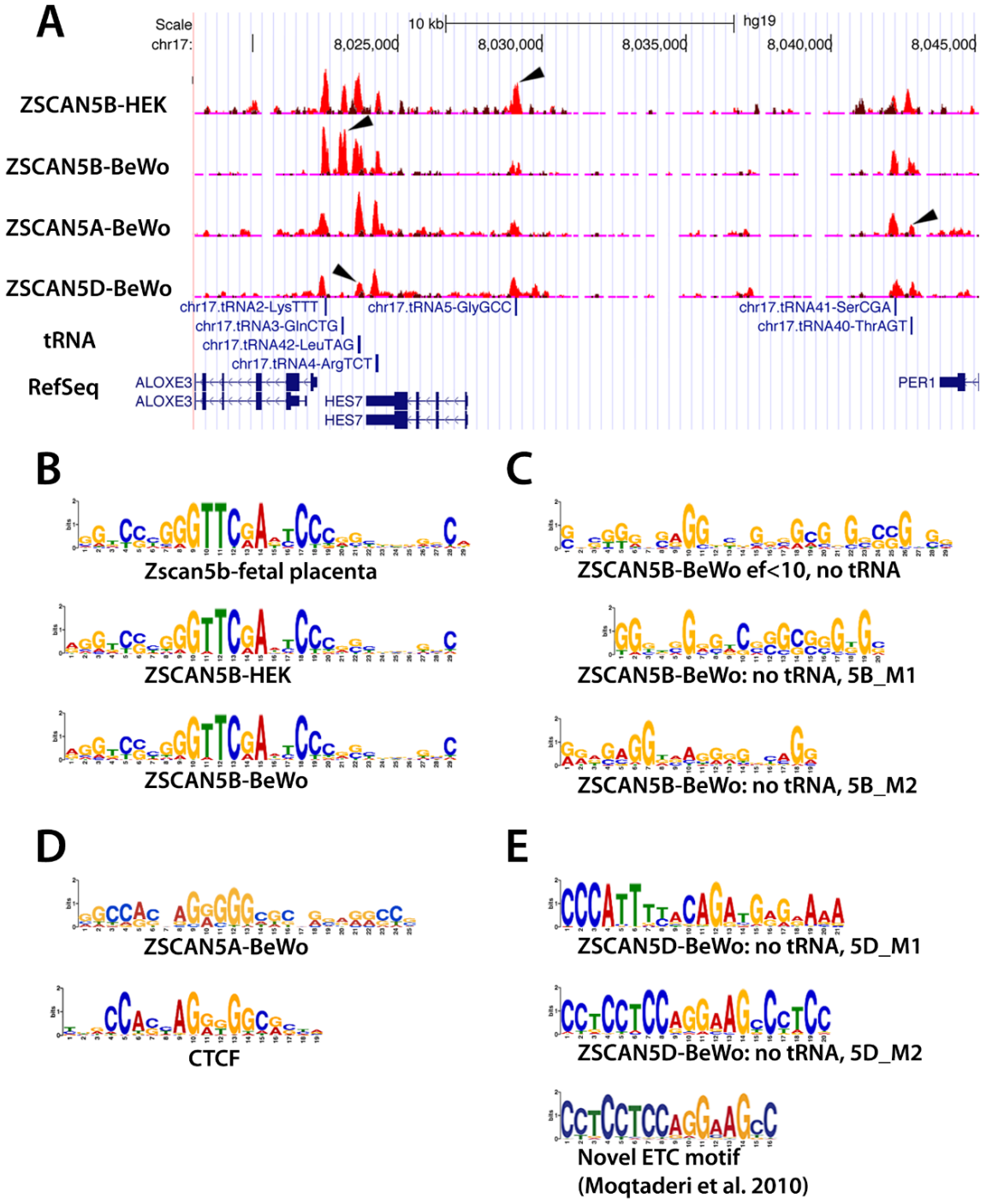

A
Scale
chr17:
10 kb
hg19
8,025,000
8,030,000
8,035,000
8,040,000
8,045,000
ZSCAN5B-HEK
ZSCAN5B-BeWo
ZSCAN5A-BeWo
ZSCAN5D-BeWo
chr17.tRNA2-LysTTT
chr17.tRNA5-GlyGCC
chr17.tRNA41-SerCGA
chr17.tRNA3-GlnCTG
chr17.tRNA40-ThrAGT
tRNA
chr17.tRNA42-LeuTAG
chr17.tRNA4-ArgTCT
RefSeq
ALOXE3
ALOXE3
HES7
HES7
PER1
B
Zscan5b-fetal placenta
ZSCAN5B-HEK
ZSCAN5B-BeWo
C
ZSCAN5B-BeWo ef<10, no tRNA
ZSCAN5B-BeWo: no tRNA, 5B_M1
ZSCAN5B-BeWo: no tRNA, 5B_M2
D
ZSCAN5A-BeWo
CTCF
E
ZSCAN5D-BeWo: no tRNA, 5D_M1
ZSCAN5D-BeWo: no tRNA, 5D_M2
Novel ETC motif
(Moqtaderi et al. 2010)

Figure 2.5. (cont.) ZSCAN5 binding displays a protein- and cell type-specific preference for tDNAs and other Pol III-related loci enriched in G/C rich motifs
(A) Distribution of ChIP-seq read peaks in the *HES7* region are displayed in a snapshot from the UCSC browser. ChIP read pileups are shown in red, with the distribution of background reads from genomic input are displayed in brown. Peaks from ZSCAN5B in chromatin from two cell lines, HEK-293 and BeWo, shown together with peak profiles from ZSCAN5A and ZSCAN5D ChIP in BeWo chromatin, reveal strong enrichment over tRNA genes that are clustered in the region; arrows in peak tracks highlight peaks that are differentially enriched by ZSCAN5B ChIP in the two cell types, or in ChIP with ZSCAN5B compared to ZSCAN5A or ZSCAN5D. (B) Motifs detected as most highly enriched and central in ChIP peaks for mouse Zscan5b and human ZSCAN5B in HEK or BeWo cells include the TFIIIC-binding B-box, which is present and highly conserved in all expressed tDNAs, and G/C-rich surrounding DNA. (C) A motif search conducted after removing tDNA sequences from the ZSCAN5B BeWo peak set revealed G/C-rich sequences, including an extended motif as well as two shorter motifs (5B_M1, 5B_M2) as predicted ZSCAN5B binding motifs. (D) A G/C rich central motif was also detected in analysis of ZSCAN5A ChIP peaks; a portion of this motif bears striking resemblance to the known motif for transcription factor and insulator protein, CTCF. (E) A distinct set of motifs were detected in high-scoring ZSCAN5D peaks, including a G/C rich motif, 5D_M2, bearing strong similarity to a novel ETC motif detected by Moqtaderi and colleagues (Moqtaderi et al., 2010).

Figure 2.6.

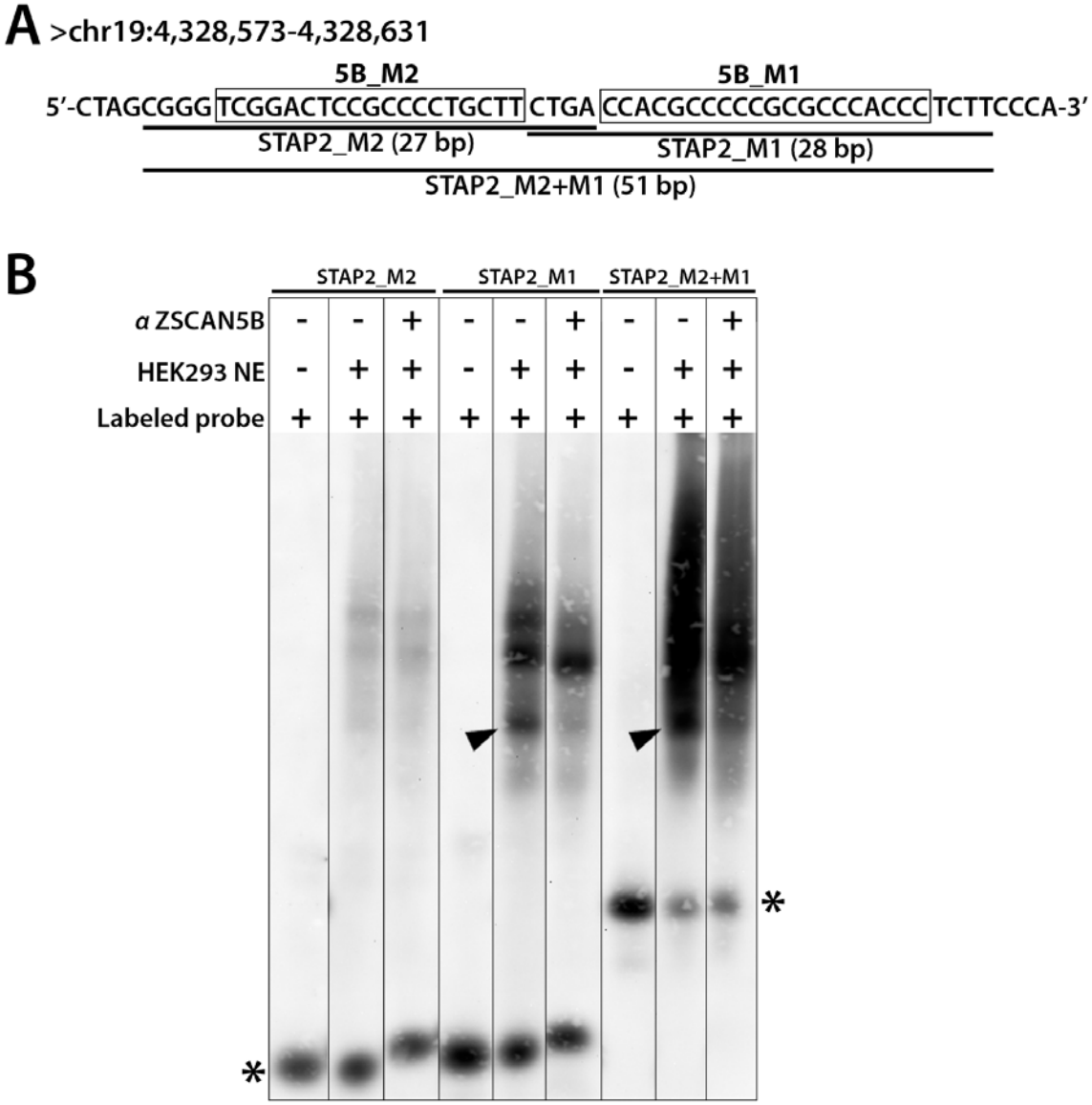

A
>chr19:4,328,573-4,328,631
5B_M2
5B_M1
5'-CTAGCGGG TCGGACTCCGCCCCTGCTT CTGA CCACGCCCCCGCGCCCACCC TCTTCCCA-3'
STAP2_M2 (27 bp)
STAP2_M1 (28 bp)
STAP2_M2+M1 (51 bp)
B
STAP2_M2
STAP2_M1
STAP2_M2+M1
α ZSCAN5B
HEK293 NE
Labeled probe
*
*

Figure 2.6. (cont.) ZSCAN5B proteins bind G/C rich motifs

To confirm that the ZSCAN5B protein binds to the G-rich motifs, we tested the summit region of a non-tDNA peak uniquely detected with high efficiency by ZSCAN5B, located within an intron of the STAP2 gene (human assembly GRCh37, chr19:4,328,490-4,328,689), which includes both the predicted 5B_M1 and 5B_M2. (A) Biotin-labeled oligonucleotides were designed to cover fragments of the two peak regions including the 5B_M2 (boxed, left) and 5B_M1 (boxed, right) motifs. Labeled probes containing 5B_M2 were named “STAP2_M2” (27 bp), 5B_M1 as “STAP2_M1” (28 bp), or 5B_M2 and 5B_M1, and an oligonucleotide probe containing both sequences was named “STAP2_M2+M1” (51 bp). (B) Nuclear extracts (NE) to which a biotinylated probe (Labeled probe) had been added for “shift”, with or without the anti-ZSCAN5B antibody (αZSCAN5B) for “supershift” were resolved on 6% non-denaturing polyacrylamide gels. Addition of αZSCAN5B caused the effective supershift of the lowest band of three “shift” complexes for STAP2_M1 and STAP2_M2+M1 (arrowheads). Asterisks indicate unbound biotinylated DNA probes.

Figure 2.7.

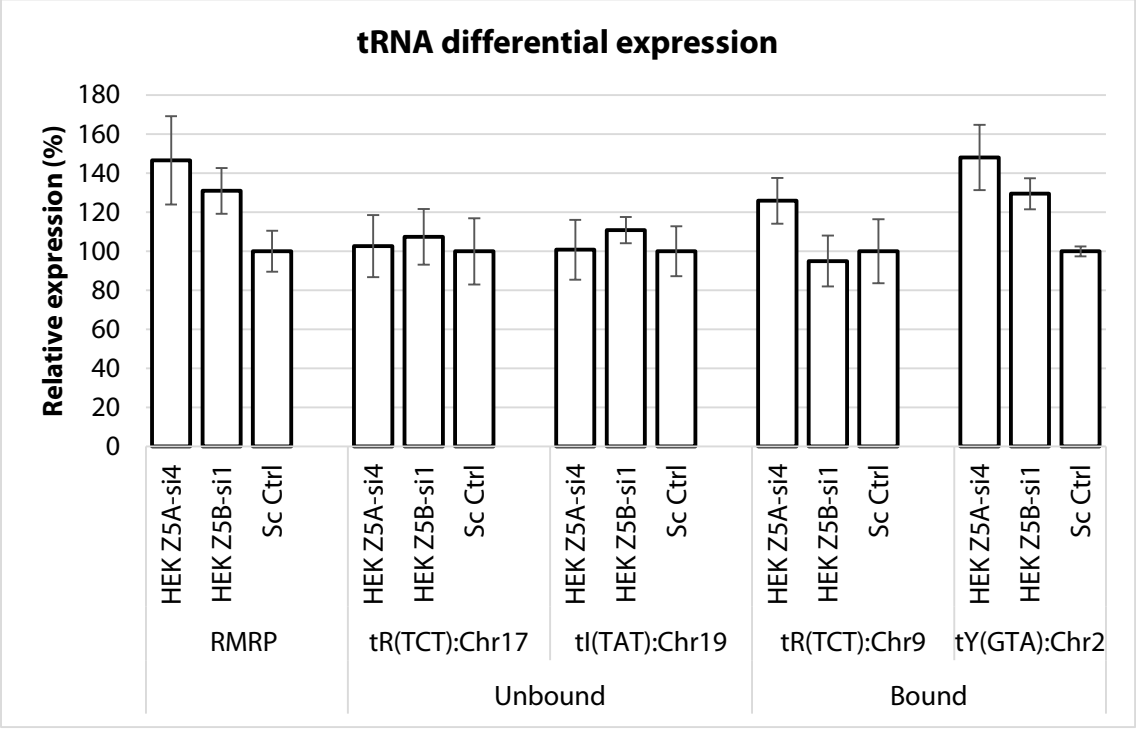

tRNA differential expression
Relative expression (%)
180
160
140
120
100
80
60
40
20
0
HEK Z5A-si4
HEK Z5B-si1
Sc Ctrl
RMRP
tR(TCT):Chr17
tI(TAT):Chr19
tR(TCT):Chr9
tY(GTA):Chr2
Unbound
Bound

Figure 2.7. (cont.) Differential expression of ZSCAN5-bound RMRP and tRNAs
qRT-PCR was done to measure the differential expression of gene loci that are bound and unbound by ZSCAN5A and ZSCAN5B after ablation of each gene by siRNA treatments in HEK-293 cells. Error bars correspond to the standard deviations between experimental triplicates.

TABLES

Table 2.1. Gene Ontology (GO) clusters identified as significantly enriched in gene sets up- or down-regulated after ZSCAN5 gene siRNA knockdown

[1] Calculated as the geometric mean (in – log scale) of *P* values for clustered terms, as described in Huang *et al.*, 2009. An enrichment of 1.3 corresponds roughly to a combined *P* value of 0.05. The top-scoring term for each cluster is reported. [2, 3] Clusters associated with Up- or Down-regulated genes, respectively.

DAVID Functional cluster	ZSCAN5A (DAVID enrichment factor) [1]		ZSCAN5B (DAVID enrichment factor)	
	Up[2]	Down[3]	Up	Down
Ribonucleotide complex/ribosome biogenesis	9.6			
M-phase/cell cycle	5.0			
Chromosome segregation/spindle	3.1			
Ubiqutin ligase complex/unfolded protein binding	2.4			
Regulation of metaphase/anaphase transition	2.2			
Condensed chromosome/kinetochore	1.9		1.9	
M-phase of meiotic cell cycle	1.9			
RNA splicing	1.8			
Noncoding RNA metabolic process /tRNA processing			4.7	
tRNA modification/wobble uridine modification			2.1	
Mitochondrion			3.4	
Macromolecular complex assembly			2.5	
Transcription regulator activity/regulation of transcription		8.9		3.6
Embryonic morphogenesis		3.2		
Pattern specification process		2.4		
Cell matrix adhesion		2.7		1.9
Extracellular matrix organization		2.2		1.6
Positive regulation of cell migration				2.3
Morphogenesis of a branching structure		2.4		
Tube development/lung development				2.0
Kidney development				1.8
Ectodermal gut morphogenesis		2.5		
Neuron cell fate commitment		2.5		
Axon guidance				1.4
Olfactory bulb development		1.7		
Blood vessel development		2.2		
Cartilage development		1.7		
Hair cycle process/epidermis development		1.7		
Wnt signaling pathway		1.4		
hematopoietic or lymphoid organ development		1.3		

Table 2.2. Relative enrichment (+) or under-representation (-) of specific repeat families in collections of ZSCAN5, RPC155, TFIIIC, and ETC ChIP peaks

Repeat Family[1]	ZSCAN5B		ZSCAN5A		ZSCAN5D		RPC155[2]		TFIIIC[2]		ETC[2]	
	+/-	p-value	+/-	p-value	+/-	p-value	+/-	p-value	+/-	p-value	+/-	p-value
tRNA	+	0	+	0	+	0	+	0	+	0		
RNA	+	1.4E-12					+	1.5E-02				
scRNA	+	1.9E-08					+	2.9e-19	+	4.3e-02		
snRNA	+	4.8E-04					+	5.3E-14				
Alu	-	2.2E-08	-	4.3E-25	-	7.9E-70	+	0	+	0	+	0
L2	-	2.3E-04			+	7.7e-10	-	1.2E-02	+	4.2e-47	+	6.4e-70
MIR	-	3.5E-03			+	1.2e-171	+	4.1E-21	+	3.1e-03	+	5.7e-04
ERV1	-	5.0E-03	-	5.0e-04	-	2.5e-14	+	3.3E-07	-	1.8e-10	-	3.7e-06
telo			+	7.5E-08								
Simple Repeat			-	4.1E-04			-	1.9E-04	+	1.9e-03		
srpRNA			+	2.3E-02			+	1.1E-03				
Gypsy			+	2.7E-02					-	4.7E-02	+	3.4E-22
Low Complexity					+	4.4E-23	-	1.4E-06	+	1.5E-50		
LTR?					+	4.8E-10						
Deu							+	3.1E-02				
LTR							+	9.7E-03				
rRNA							+	0	+	0	+	1.8E-07
RTE-BovB							+	2.2E-03				
SINE											+	2.3E-02

[1]Repeat family names and locations taken from repeat masker, http://www.repeatmasker.org/ ; a full accounting with enrichments and depletions for specific elements in each family is provided in *Table A.4*.

[2] Peak coordinates for these feature types are taken from Moqtaderi *et al.*, 2011, and were lifted over to human genome sequence build hg19.

REFERENCES CITED

1. Barski, A., Chepelev, I., Liko, D., Cuddapah, S., Fleming, A.B., Birch, J., Cui, K., White, R.J., and Zhao, K. (2010). Pol II and its associated epigenetic marks are present at Pol III-transcribed noncoding RNA genes. Nature structural & molecular biology *17*, 629-634.

2. Berg, J.M. (1992). Sp1 and the subfamily of zinc finger proteins with guanine-rich binding sites. Proceedings of the National Academy of Sciences of the United States of America *89*, 11109-11110.

3. Canella, D., Bernasconi, D., Gilardi, F., LeMartelot, G., Migliavacca, E., Praz, V., Cousin, P., Delorenzi, M., Hernandez, N., and Cycli, X.C. (2012). A multiplicity of factors contributes to selective RNA polymerase III occupancy of a subset of RNA polymerase III genes in mouse liver. Genome research *22*, 666-680.

4. Canella, D., Praz, V., Reina, J.H., Cousin, P., and Hernandez, N. (2010). Defining the RNA polymerase III transcriptome: Genome-wide localization of the RNA polymerase III transcription machinery in human cells. Genome research *20*, 710-721.

5. Collins, T., Stone, J.R., and Williams, A.J. (2001). All in the family: the BTB/POZ, KRAB, and SCAN domains. Mol Cell Biol *21*, 3609-3615.

6. Comet, I., Schuettengruber, B., Sexton, T., and Cavalli, G. (2011). A chromatin insulator driving three-dimensional Polycomb response element (PRE) contacts and Polycomb association with the chromatin fiber. Proceedings of the National Academy of Sciences of the United States of America *108*, 2294-2299.

7. Cordaux, R., and Batzer, M.A. (2009). The impact of retrotransposons on human genome evolution. Nature reviews. Genetics *10*, 691-703.

8. Crepaldi, L., Policarpi, C., Coatti, A., Sherlock, W.T., Jongbloets, B.C., Down, T.A., and Riccio, A. (2013). Binding of TFIIIC to sine elements controls the relocation of activity-dependent neuronal genes to transcription factories. PLoS genetics *9*, e1003699.

9. Cuddapah, S., Jothi, R., Schones, D.E., Roh, T.Y., Cui, K., and Zhao, K. (2009). Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. Genome research *19*, 24-32.

10. Dittmar, K.A., Goodenbour, J.M., and Pan, T. (2006). Tissue-specific differences in human transfer RNA expression. PLoS genetics *2*, e221.

11. Eisenberg, E., and Levanon, E.Y. (2003). Human housekeeping genes are compact. Trends in genetics : TIG *19*, 362-365.

12. Felton-Edkins, Z.A., Fairley, J.A., Graham, E.L., Johnston, I.M., White, R.J., and Scott, P.H. (2003). The mitogen-activated protein (MAP) kinase ERK induces tRNA synthesis by phosphorylating TFIIIB. EMBO J *22*, 2422-2432.

13. Gilbert, N., and Labuda, D. (2000). Evolutionary inventions and continuity of CORE-SINEs in mammals. Journal of molecular biology *298*, 365-377.

14. Gill, T., Cai, T., Aulds, J., Wierzbicki, S., and Schmitt, M.E. (2004). RNase MRP cleaves the CLB2 mRNA to promote cell cycle progression: novel method of mRNA degradation. Mol Cell Biol *24*, 945-953.

15. Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C.K., Chrast, J., Lagarde, J., Gilbert, J.G., Storey, R., Swarbreck, D., *et al.* (2006). GENCODE: producing a reference annotation for ENCODE. Genome biology *7 Suppl 1*, S4 1-9.

16. Hiraga, S., Botsios, S., Donze, D., and Donaldson, A.D. (2012). TFIIIC localizes budding yeast ETC sites to the nuclear periphery. Molecular biology of the cell *23*, 2741-2754.

17. Huang, D.W., Sherman, B.T., Tan, Q., Collins, J.R., Alvord, W.G., Roayaei, J., Stephens, R., Baseler, M.W., Lane, H.C., and Lempicki, R.A. (2007a). The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome biology *8*, R183.

18. Huang, D.W., Sherman, B.T., Tan, Q., Kir, J., Liu, D., Bryant, D., Guo, Y., Stephens, R., Baseler, M.W., Lane, H.C., *et al.* (2007b). DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic acids research *35*, W169-175.

19. Hull, M.W., Erickson, J., Johnston, M., and Engelke, D.R. (1994). tRNA genes as transcriptional repressor elements. Mol Cell Biol *14*, 1266-1277.

20. Jiang, H., and Peterlin, B.M. (2008). Differential chromatin looping regulates CD4 expression in immature thymocytes. Mol Cell Biol *28*, 907-912.

21. Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J. (2004). The UCSC Table Browser data retrieval tool. Nucleic acids research *32*, D493-496.

22. Kim, J., Bergmann, A., Wehri, E., Lu, X., and Stubbs, L. (2001). Imprinting and evolution of two Kruppel-type zinc-finger genes, ZIM3 and ZNF264, located in the PEG3/USP29 imprinted domain. Genomics *77*, 91-98.

23. Kim, J., Kollhoff, A., Bergmann, A., and Stubbs, L. (2003). Methylation-sensitive binding of transcription factor YY1 to an insulator sequence within the paternally expressed imprinted gene, Peg3. Hum. Mol. Genet. *12*, 233-245.

24. Kinsey, P.T., and Sandmeyer, S.B. (1991). Adjacent pol II and pol III promoters: transcription of the yeast retrotransposon Ty3 and a target tRNA gene. Nucleic acids research *19*, 1317-1324.

25. Kirchner, S., and Ignatova, Z. (2015). Emerging roles of tRNA in adaptive translation, signalling dynamics and disease. Nature reviews. Genetics *16*, 98-112.

26. Kirkland, J.G., Raab, J.R., and Kamakaka, R.T. (2013). TFIIIC bound DNA elements in nuclear organization and insulation. Biochimica et biophysica acta *1829*, 418-424.

27. Kuhn, R.M., Haussler, D., and Kent, W.J. (2013). The UCSC genome browser and associated tools. Briefings in bioinformatics *14*, 144-161.

28. Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology *10*, R25.

29. Lee, Y.L., Li, Y.C., Su, C.H., Chiao, C.H., Lin, I.H., and Hsu, M.T. (2015). MAF1 represses CDKN1A through a Pol III-dependent mechanism. eLife *4*.

30. Levine, M., Cattoglio, C., and Tjian, R. (2014). Looping back to leap forward: transcription enters a new era. Cell *157*, 13-25.

31. Li, T., Hu, J.F., Qiu, X., Ling, J., Chen, H., Wang, S., Hou, A., Vu, T.H., and Hoffman, A.R. (2008). CTCF regulates allelic expression of Igf2 by orchestrating a promoter-polycomb repressive complex 2 intrachromosomal loop. Mol Cell Biol *28*, 6473-6482.

32. Liu, H., Chang, L.H., Sun, Y., Lu, X., and Stubbs, L. (2014). Deep vertebrate roots for mammalian zinc finger transcription factor subfamilies. Genome biology and evolution *6*, 510-525.

33. Machanick, P., and Bailey, T.L. (2011). MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics *27*, 1696-1697.

34. Martin, A.N., and Li, Y. (2007). RNase MRP RNA and human genetic diseases. Cell research *17*, 219-226.

35. McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nature biotechnology *28*, 495-501.

36. Moqtaderi, Z., Wang, J., Raha, D., White, R.J., Snyder, M., Weng, Z., and Struhl, K. (2010). Genomic binding profiles of functionally distinct RNA polymerase III

transcription complexes in human cells. Nature structural & molecular biology *17*, 635-640.

37. Noma, K., Cam, H.P., Maraia, R.J., and Grewal, S.I. (2006). A role for TFIIIC transcription factor complex in genome organization. Cell *125*, 859-872.

38. Oler, A.J., Alla, R.K., Roberts, D.N., Wong, A., Hollenhorst, P.C., Chandler, K.J., Cassiday, P.A., Nelson, C.A., Hagedorn, C.H., Graves, B.J., *et al.* (2010). Human RNA polymerase III transcriptomes and relationships to Pol II promoter chromatin and enhancer-binding factors. Nature structural & molecular biology *17*, 620-628.

39. Pascali, C., and Teichmann, M. (2013). RNA polymerase III transcription - regulated by chromatin structure and regulator of nuclear chromatin organization. Sub-cellular biochemistry *61*, 261-287.

40. Pavletich, N.P., and Pabo, C.O. (1991). Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. Science *252*, 809-817.

41. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

42. Raab, J.R., Chiu, J., Zhu, J., Katzman, S., Kurukuti, S., Wade, P.A., Haussler, D., and Kamakaka, R.T. (2012). Human tRNA genes function as chromatin insulators. EMBO J *31*, 330-350.

43. Raha, D., Wang, Z., Moqtaderi, Z., Wu, L., Zhong, G., Gerstein, M., Struhl, K., and Snyder, M. (2010). Close association of RNA polymerase II and many transcription factors with Pol III genes. Proceedings of the National Academy of Sciences of the United States of America *107*, 3639-3644.

44. Rideout, E.J., Marshall, L., and Grewal, S.S. (2012). Drosophila RNA polymerase III repressor Maf1 controls body size and developmental timing by modulating tRNAiMet synthesis and systemic insulin signaling. Proceedings of the National Academy of Sciences of the United States of America *109*, 1139-1144.

45. Rogler, L.E., Kosmyna, B., Moskowitz, D., Bebawee, R., Rahimzadeh, J., Kutchko, K., Laederach, A., Notarangelo, L.D., Giliani, S., Bouhassira, E., *et al.* (2014). Small RNAs derived from lncRNA RNase MRP have gene-silencing activity relevant to human cartilage-hair hypoplasia. Human molecular genetics *23*, 368-382.

46. Schmitt, B.M., Rudolph, K.L., Karagianni, P., Fonseca, N.A., White, R.J., Talianidis, I., Odom, D.T., Marioni, J.C., and Kutter, C. (2014). High-resolution mapping of transcriptional dynamics across tissue development reveals a stable mRNA-tRNA interface. Genome research *24*, 1797-1807.

47. Schneider, C.A., Rasband, W.S., and Eliceiri, K.W. (2012). NIH Image to ImageJ: 25 years of image analysis. Nat methods *9*, 671-675.

48. Schumacher, C., Wang, H., Honer, C., Ding, W., Koehn, J., Lawrence, Q., Coulis, C.M., Wang, L.L., Ballinger, D., Bowen, B.R., *et al.* (2000). The SCAN domain mediates selective oligomerization. J Biol Chem *275*, 17173-17179.

49. Shaw, G., Morse, S., Ararat, M., and Graham, F.L. (2002). Preferential transformation of human neuronal cells by human adenoviruses and the origin of HEK 293 cells. FASEB journal : official publication of the Federation of American Societies for Experimental Biology *16*, 869-871.

50. Stutz, F., Gouilloud, E., and Clarkson, S.G. (1989). Oocyte and somatic tyrosine tRNA genes in Xenopus laevis. Genes & development *3*, 1190-1198.

51. Thiel, C.T., Mortier, G., Kaitila, I., Reis, A., and Rauch, A. (2007). Type and level of RMRP functional impairment predicts phenotype in the cartilage hair hypoplasia-anauxetic dysplasia spectrum. American journal of human genetics *81*, 519-529.

52. Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols *7*, 562-578.

53. Ullu, E., and Tschudi, C. (1984). Alu sequences are processed 7SL RNA genes. Nature *312*, 171-172.

54. Van Bortle, K., and Corces, V.G. (2012). tDNA insulators and the emerging role of TFIIIC in genome organization. Transcription *3*, 277-284.

55. Van Bortle, K., Nichols, M.H., Li, L., Ong, C.T., Takenaka, N., Qin, Z.S., and Corces, V.G. (2014). Insulator function and topological domain border strength scale with architectural protein occupancy. Genome biology *15*, R82.

56. Wang, Q., Nowak, C.M., Korde, A., Oh, D.H., Dassanayake, M., and Donze, D. (2014). Compromised RNA polymerase III complex assembly leads to local alterations of intergenic RNA polymerase II transcription in Saccharomyces cerevisiae. BMC biology *12*, 89.

57. Whitfield, M.L., Sherlock, G., Saldanha, A.J., Murray, J.I., Ball, C.A., Alexander, K.E., Matese, J.C., Perou, C.M., Hurt, M.M., Brown, P.O., *et al.* (2002). Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Molecular biology of the cell *13*, 1977-2000.

58. Wolfe, S.A., Grant, R.A., Elrod-Erickson, M., and Pabo, C.O. (2001). Beyond the "recognition code": structures of two Cys2His2 zinc finger/TATA box complexes. Structure *9*, 717-723.

59. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., *et al.* (2008). Model-based analysis of ChIP-Seq (MACS). Genome biology *9*, R137.

CHAPTER 3:

A ROLE FOR MAMMALIAN SCAN-CONTAINING ZINC FINGER TRANSCRIPTION FACTOR ZSCAN5A IN CELL CYCLE PROGRESSION

Younguk Sun[1,2], and Lisa Stubbs[1,2,3]

ABSTRACT

RNA Polymerase III (Pol III) transcription is critical for cell survival. In addition to its canonical roles in tRNA and ribosomal RNA synthesis, Pol III transcription is essential to a variety of other nuclear processes, including nucleosome positioning, global genome and subnuclear organization, and has direct effects on RNA pol II transcription. We have previously identified the ZSCAN5 transcription factor family – including conserved mammalian *ZSCAN5B* and primate-specific paralogs, *ZSCAN5A* and *ZSCAN5D* – as proteins that occupy Pol III promoters and extra-TFIIIC (ETC) sites throughout the genome. ZSCAN5 protein binding regulates expression of tRNA and other Pol III genes, but also affects the expression of neighboring Pol II genes involved in ribosome biogenesis, tRNA processing, cell cycle progression and cellular differentiation. Here we demonstrate that human ZSCAN5 genes are cell cycle regulated with a similar peak of expression around the time of mitosis, and show that the proteins display strong and distinct patterns of cellular localization in the dividing cells. Together with previous results, these data suggested an important role in chromosome segregation and mitotic progression, especially for *ZSCAN5A*. To test this hypothesis, we manipulated the levels of *ZSCAN5A* expression using stably integrated, tetracycline regulated constructs in cultured human cells. Consistent with an essential role in cell cycle progression *ZSCAN5A* knockdown led to the accumulation of cells in the mitotic phase and the appearance of aneuploid cells. These data suggest that *ZSCAN5A* has adopted an essential role in chromatin condensation and segregation in human cells, through combined action on Pol III transcripts and position effects on essential Pol II genes.

[1] Institute for Genomic Biology

[2] Department of Cell and Developmental Biology

University of Illinois at Urbana-Champaign, Urbana IL 61801

[3] Corresponding author

Keywords:

Zinc finger transcription factor/cell cycle progression/RNA Polymerase III transcription/tRNA/chromatin architecture

INTRODUCTION

tRNAs are encoded by RNA Polymerase III-transcribed genes that are randomly dispersed throughout the genomes in all eukaryotes, from the yeasts, *S. cerevisiae* and *S. pombe*, to mammals (Canella et al., 2010; Dittmar et al., 2006; Harismendy et al., 2003; Moqtaderi and Struhl, 2004; Roberts et al., 2003) . While the primary function of tRNA genes (tDNA) is to provide templates for the transcription of essential tRNA molecules, a number of studies have reported that these genes also serve a variety of "extra-transcriptional" functions (Ebersole et al., 2011; Hull et al., 1994; Kendall et al., 2000; Kinsey and Sandmeyer, 1991; Raab et al., 2012). These functions have been most extensively investigated in yeasts, where the genes are known to congregate together at the nucleolus and centromeres to perform non-canonical functions (Thompson et al., 2003; Wang et al., 2005).

One of the most interesting roles for tDNA is that it can serve as a heterochromatin barrier, that is, to prevent spreading of heterochromatin to euchromatic regions by repositioning nucleosomes or recruiting chromatin remodeling complex such as histone acetyltransferases (HATs) (Bachman et al., 2005; Donze et al., 1999; Donze and Kamakaka, 2001; Gelbart et al., 2005; Mertens and Roeder, 2008). Loss of this function was reported to result in abnormal meiotic chromosome segregation in *S. pombe* (Gaither et al., 2014) and in *S. cerevisiae*, suppression or deletion of tDNA loci is associated with abnormal expression of nearby Polymerase 2 (Pol II) genes (Haeusler et al., 2008; Wang et al., 2005). Pol III transcripts, and derivative elements called extra-TFIIIC (ETC) sites because they bind the Pol III transcription factor TFIIIC, have also been demonstrated to serve as insulators and chromatin organizing elements in mammalian genomes (Moqtaderi et al., 2010; Oler et al., 2010; Raab et al., 2012). Cohesin-interacting CTCF is also known to be enriched at ETCs and tDNAs with high TFIIIC and Pol III enrichment (Moqtaderi et al., 2010; Oler et al., 2010).

The expression of tRNA is regulated under many conditions, including during cell cycle progression and cell proliferation (Marshall and White, 2008). Most studies of Pol III regulation have concluded that control is mediated through TFIIIB (Marshall and White, 2008; White et al., 1995). It is also worth noting that 1) the relative proportions of

individual tRNA vary greatly in different cell types (Dittmar et al., 2006) and 2) there is limited overlap between the tDNAs that are occupied with Pol III in different cell types and tissues (Barski et al., 2010). These findings contradict previous simplistic models of Pol III regulation, which suggested that the genes were all expressed together, and suggest a more complex regulation involving a host of different TFs binding to the individual loci.

In our previous study, we employed chromatin immunoprecipitation followed by deep sequencing (ChIP-seq) to reveal genome-wide binding of the of ZSCAN5 transcription factor (TF) family to Pol III promoters and ETC sites (Sun et al., 2016). This small TF subfamily was founded by a unique, conserved member, *ZSCAN5B,* which duplicated specifically in early primate history to generate three additional TF genes. While all ZSCAN5 proteins bind tDNAs, ZSCAN5B exhibited the most strikingly high binding specificity to tRNA genes, whereas primate-specific ZSCAN5A, which is the dominantly expressed member of the primate family, also preferentially bound to tDNAs but also more widely at Pol III-related transposable elements (TEs). Furthermore, another primate duplicate, ZSCAN5D, displayed a high preference for a subset of TE-derived ETC sites. Transcriptome analysis after ZSCAN5 gene "knockdown" revealed that ZSCAN5A and ZSCAN5B not only regulate the expression of tRNA and other Pol III genes, but also modulates the expression of neighboring Pol II genes. Based on these data we hypothesized that ZSCAN5A and ZSCAN5B cooperatively interact in human cells, and that ZSCAN5A plays a particularly important role in the regulation of mitotic progression. This last inference was particularly intriguing, since *ZSCAN5A* was previously identified in a previous study as one of ~850 genes expressed in a cell cycle stage-specific pattern, with transcription peaking around the M/G1 transition (Whitfield et al., 2002).

The goal of the present study was to further investigate the potential roles of ZSCAN5 genes in regulating cell cycle progression with a special focus on *ZSCAN5A.* Consistent with functional predictions, we show that the human ZSCAN5 paralogs are cell-cycle regulated with peak expression around the time of mitosis. Using immunocytochemistry with antibodies specific to ZSCAN5A, ZSCAN5B, and ZSCAN5D, we demonstrate that the proteins display distinct patterns of cellular

localization that is significantly enhanced in dividing cells. A series of human cell lines engineered to stably carry a tetracycline-inducible transgenic allowed the manipulation of ZSCAN5A expression, and revealed that ZSCAN5A depletion leads to the dysregulation of cell cycle. More specifically, ZSCAN5A depletion leads to the accumulation of cells in the mitotic (M) phase and the appearance of aneuploid cells.

These data suggest that, either through a cyto-architectural function, regulation of Pol III transcription, position effects on Pol II genes, or a combination of these roles, primate-specific ZSCAN5A has acquired a critical role in an ancient function, that is, maintenance of genome integrity in human cells.

RESULTS

Human ZSCAN5 genes display a peak of mitotic expression

As mentioned above, *ZSCAN5A* was identified in a previous study as one of ~850 genes expressed in a cell cycle stage-specific pattern, with transcription peaking around the M/G1 transition (Whitfield et al., 2002). To confirm this finding and to investigate the cell-cycle expression patterns for other *ZSCAN5* genes, we synchronized HEK-293 cultures using a double thymidine (TT) block, which synchronizes the cells at early S-phase (Bostock et al., 1971). We collected cells at different time points after release to test gene expression levels of *ZSCAN5A*, *ZSCAN5B*, and *ZSCAN5D* by qRT-PCR, together with marker genes expressed at specific cell cycle stages [*CCNE1* (expression peak at G1/S), *RRM2* (S), *CDC2* (G2), *BUB1* (G2/M) and *PTTG1* (M/G1)]. *ZSCAN5B* and *ZSCAN5D* showed clear peaks of expression beginning around 14 hr after synchronization, consistent with peak transcription during the M/G1 transition (*Fig. 1*). *ZSCAN5A* expression showed a similar expression pattern consistent with the published reports although RNA levels peaked somewhat sooner than the M/G1 transition, beginning around G2/M in our experiments.

ZSCAN5s are differentially localized during the cell cycle

In yeasts, tDNA and other Pol III transcripts concentrate within subnuclear structures, including the nucleolus (Thompson et al., 2003; Wang et al., 2005) or the

mitotic spindle and centrosome (Snider et al., 2014) in budding yeast, and the nuclear periphery in fission yeast (Huang and Maraia, 2001). In both species, tDNAs have been demonstrated to interact with cohesin, which plays an important role in chromosome tethering throughout the cell cycle. No such distinct localization has been reported for tDNA or other Pol III transcripts in mammalian cells. The cellular locations of the ZSCAN5 proteins was therefore of interest.

To examine the localization of ZSCAN5 proteins in mitotic cells, we used tested antibodies (Sun et al., 2016) along with a centrosomal marker (anti-gamma tubulin antibody) for immunocytochemistry. Our previous study showed robust expression of ZSCAN5 proteins in HEK-293 cells, and therefore we used the same cell line cultured in asynchronous manner to capture the cell cycle stage-specific localization of the ZSCAN5 proteins.

While all three ZSCAN5 proteins were detected throughout the cell cycle, a few noticeable localization patterns were observed. For example, during interphase, ZSCAN5A was concentrated in the nuclei of either recently divided cells or cells ready for mitotic division, indicating that ZSCAN5A expression is more pronounced during late G2/M transition and early G1, concordant with gene expression (*Fig 3.1*, *Fig 3.2A*). However, more striking concentration of the protein was found during M-phase, and especially anaphase. Although ZSCAN5A protein was widely distributed and not restricted to certain subcellular organelles during mitosis (*Fig. 3.2B*), ZSCAN5B was concentrated particularly in the centrosomal region during early to late anaphase (*Fig. 3.2C*). In contrast, ZSCAN5D protein was concentrated in close proximity to the cleavage furrow in the late anaphase to early telophase (*Fig. 3.2D*). This differential cellular localization is remarkable as all three ZSCAN5 members exhibit similar cell cycle stage-specific expression (*Fig. 3.1*) and all three proteins bind to RNA Pol III loci (Sun et al. 2016). These data suggest that mammalian ZSCAN5B and its primate-specific paralogs may have acquired novel functions involved in cell division, and more specifically during mitotic division, despite their similar expression patterns and binding sites.

Notably since ZSCAN5B displays a compelling preference for tDNA loci (Sun et al., 2016), its centrosomal co-localization is particularly interesting given previous

studies that have reported that tDNA co-localize with the centrosome in budding yeast (Snider et al., 2014). These data suggest that tDNAs are also concentrated at centrosomes during mitosis in mammalian cells, the first time that such a conserved localization has been noted in any study.

ZSCAN5A knockdown increases numbers of M-phase cells and aneuploidy

Consistent with previous reports (Whitfield et al., 2002), our data demonstrate that ZSCAN5 paralogs are commonly up-regulation around the time of mitosis (*Fig. 3.1*). Adding to the intrigue, *ZSCAN5A* knockdown experiments strongly suggested a role for this gene in the regulation of mitotic spindle formation, chromosome segregation and the metaphase/anaphase transition (Sun et al. 2016). To test this hypothesis, we transfected HEK-293 cells with plasmid constructs designed to express a tetracycline-regulated (Tet-on) short hairpin RNA (shRNA), with the goal of knocking down *ZSCAN5A* controllably and stably in HEK-293 cells (ZSCAN5A-Tet-shRNA). The transfected, doxycycline treated cells showed significant knockdown of *ZSCAN5A*, as confirmed with qRT-PCR (approximately 85%, not shown).

Additionally, we created versions of cell lines that conditionally overexpress either the full-length ZSCAN5A or ZSCAN5B open reading frame to investigate the dosage-dependent cellular phenotypes (ZSCAN5A-Tet-OE and ZSCAN5B-Tet-OE). qRT-PCR validation of these overexpression constructs after induction showed about 10~25-fold overexpression of respective transcripts (not shown), Further analysis of these cells revealed phenotypes very consistent with the hypothesis that ZSCAN5A regulates cell cycle progression and in particular, chromosome segregation and metaphase-anaphase transition. In contrast to cells transfected with the empty plasmid vector, ZSCAN5A-Tet-OE, and ZSCAN5B-Tet-OE, a significantly larger fraction of *ZSCAN5A* shRNA-expressing cells were detected at M-phase, suggesting a less rapid and efficient transition to anaphase (*Fig. 3.3*). Furthermore, flow cytometric experiments revealed that the appearance of aneuploidy after *ZSCAN5A* knockdown (*Fig. 3.4*), consisting of a population of cells with DNA content close to a haploid complement (*Fig. 3.4D*). These aneuploid cells were not detected in any of the over-expressing cells or in cells containing the empty plasmid vectors. Taken together with the siRNA gene expression results, these

data suggest that *ZSCAN5A* depletion leads to abnormalities in spindle assembly or attachment during M-phase, a situation well known to cause metaphase arrest and aneuploidy in mammalian cells (Musacchio, 2015).

DISCUSSION

Data presented here show that the ZSCAN5 subfamily of human TF genes are regulated during the cell cycle, with a shared pattern of up-regulation around the time of mitotic M-phase. *ZSCAN5A* appears to be activated slightly earlier than *ZSCAN5B* and *ZSCAN5D,* first appearing around the G2 phase, and this fact may be relevant to its predicted dominant effects on cell cycle progression. Together with previously reported analysis of differential gene expression after *ZSCAN5A* siRNA knockdown (Sun et al., 2016), data presented suggest that the depletion of *ZSCAN5A* disrupts chromosome attachment or cohesion during the metaphase-anaphase transition, a situation well known to result in the generation of aneuploid cells (Decordier et al., 2008). Consistent with this interpretation, we also noted an unusual accumulation of metaphase cells after ZSCAN5A knockdown, suggesting a failure to pass the spindle assembly checkpoint (Musacchio, 2015). These results suggest that ZSCAN5A might play a more specific role in the maintenance of chromosome structure and integrity; through that role and due to its mitotic expression, the protein also impacts cell proliferation.

The exact mechanism through which ZSCAN5A might help maintain chromosome integrity and during mitosis will require further study. However, previous studies provide some interesting clues. In particular, we have previously shown that either directly or through “position effects”, ZSCAN5A regulate genes required for mitotic progression (Sun et al., 2016); the protein may thus function simply as the transcriptional modulator of those essential genes. However, published data on other ZNF proteins also suggest that ZSCAN5A function might be more complex. For example, *TRPS1,* the gene mutated in human Tricho-Rhino-Phalangeal Syndrome, encodes a multi zinc finger nuclear regulator of chondrocyte proliferation and differentiation (Wuelling et al., 2013). Very much like the phenotype we observed after ZSCAN5A knockdown, the loss of *Trps1* in mice led to an increased proportion of cells arrested in mitosis and

subsequently, to chromosome segregation defects (Wuelling et al., 2013). The Trps1 protein was found to interact with Hdac1 and Hdac4 to increase histone deacetylation activity, and loss of Trps1 results in histone H3 hyperacetylation which is maintained during mitosis, causing impaired chromatin condensation and HP1 binding, to have cells accumulated in prometaphase (Wuelling et al., 2013).

These possibilities may be addressed in the future if the binding partners of ZSCAN5A in various cell contexts can be identified. For example, interactions between ZSCAN5A and HDACs or other predicted partners in different cell-cycle stages can be tested individually with co-immunoprecipitation methods, or more globally with mass spectrometry-based methods (Free et al., 2009). Given the known interaction of TFIIIC and cohesin and the importance of cohesin to maintenance of chromosome structure (Hakimi et al., 2002; Parelho et al., 2008), interactions between ZSCAN5 proteins and cohesin components could be particularly enlightening. Here is it interesting to note that the differentially expressed genes detected after *ZSCAN5A* knockdown included two loci encoding essential cohesion components, *STAG2* and *SMC3* (Sun et al., 2016), making these particularly interesting subjects for future study. These experiments will be major goals of further projects centered on the ZSCAN5 family genes.

Together these data define unexpected functions for a conserved mammalian gene and two paralogs found only in primate genomes. The human paralogs have evolved from the basic tRNA-binding function of the conserved founder gene, *ZSCAN5B,* to bind more widely to Pol III transcripts and other TFIIIC binding sites and affect the transcription of a wider selection of nearby genes. Data presented here indicate that despite being duplicated only in primates, *ZSCAN5A* has been integrated into the control of a very ancient process, namely chromosome segregation during mitosis, and plays an important role in chromosome integrity. The new tools and resources I have developed (See Chapter 4) provide essential tools that will be needed to explore those functions in further depth.

MATERIALS AND METHODS

RNA PREPARATION AND QUANTITATIVE RT-PCR

Total RNA was isolated from cell lines and tissues using TRIzol (Invitrogen) followed by 30 minutes of RNase-free DNaseI treatment (New England Biolabs) at 37ºC and RNA Clean & Concentrator^{TM}-5 (Zymo Research). 2 µg of total RNA was used to generate cDNA using Superscript III Reverse Transcriptase (Invitrogen) with random hexamers (Invitrogen) according to manufacturer's instructions.

The resulting cDNAs were analyzed of transcript-specific expression through quantitative reverse-transcript PCR (qRT-PCR) using Power SYBR Green PCR master mix (Applied Biosystems, you are supposed to say where the companies are) with custom-designed primer sets (*Table 3.1*) purchased from Integrated DNA Technology. Relative expression was determined by normalizing the expression of all genes of interest to either human or mouse Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, zeta polypeptide (*YWHAZ*) expression (ΔCt) as described (Eisenberg and Levanon, 2003).

CELL CULTURE AND CELL CYCLE SYNCHRONIZATION

HEK-293 (ATCC, CRL-1573) and HEK-293 Tet-on (Clontech) cells were cultured in Dulbecco's Modified Eagle's Medium (DMEM) containing 2 mM L-glutamine, 10% Tetracycline-free fetal bovine serum (FBS, Clontech), 1X Pen Strep on 0.2% gelatin-coated flasks and incubated at 37 °C in 5% CO2. For HEK-293 cell cycle synchronization, 2 mM thymidine was added to HEK-293 cells grown to about 30% confluency and the cells were subsequently incubated for 18 hours. Thymidine was removed by washing with 1X PBS three times, and adding fresh media, followed by a further 9 hour- incubation. The second round of 2 mM thymidine was then added and cells were incubated for an additional 15 hours. Cells were released from G1/S to S by washing with 1X PBS and adding fresh media, and were collected at different time points thereafter (Bostock et al., 1971).

PLASMIDS AND TRANSFECTIONS

To establish inducible cell lines that over-express full-length ZSCAN5A, ZSCAN5B, we cloned the full open reading frames of ZSCAN5A (NM_024303) and ZSCAN5B (NM_001080456) into Tet-inducible pTRE-Tight expression vector (Clontech), and named as ZSCAN5A-Tet-OE and ZSCAN5B-Tet-OE (*Fig. A.1* and *A.2*). In addition, to create an inducible cell line that expresses short hairpin RNAs (shRNAs) targeting ZSCAN5A, annealed double-stranded oligonucleotides that encode shRNA sequence was cloned into the pSuperior.Puro plasmid (Oligoengine) and named as ZSCAN5A-Tet-shRNA (*Fig. A.3*). All the primers and oligonucleotides used in this study are described in *Table 3.1*. Resulting plasmids were prepared using a Plasmid Midi Prep kit (Qiagen) and sequenced before subsequent transfection. For plasmid DNA transfection, about $4.5x10^5$ HEK-293 cells were seeded to 6-well plates 24 hours before, and 3 µg of each plasmid; ZSCAN5A-OE or ZSCAN5B-OE, or blank plasmid along with 300 ng of pPur plasmid (Clontech) which confers puromycin resistance, or ZSCAN5A-shRNA was transfected using Lipofectamine 2000 (Invitrogen). 24 hours later, transfected cells were selected under 1 µg/ml of puromycin for additional 14 days. Single colonies were selected and expanded, then tested for the efficiency of ZSCAN5A by qRT-PCR knockdown 48 h after addition of 1 µg/mL doxycycline (Dox; Sigma-Aldrich). Each colony was analyzed of it overexpression or knockdown of respective genes and proteins compared to Dox-treated cells carrying the empty vector.

IMMUNOCYTOCHEMISTRY AND M-PHASE COUNTING

Stably transfected cell lines were maintained in a regular medium described above with the addition of 0.5 µg/ml puromycin. For immunocytochemistry (ICC), each cell was seeded and grown on gelatin-coated cover slips for 24 hours using the condition described above with or without 1 µg/ml of doxycycline. After washing with 1X PBS, cells were fixed with cold methanol and incubated at -20^{o}C for 30 mins and wash three time with 1X PBS again. Blocking was done by incubating the coverslips with antibody diluent (Dako) at 4^{o}C for 2 hours. Primary antibodies targeting ZSCAN5A (1:2000 dilution) was custom generated by Abgent, while those targeting ZSCAN5B (Santa Cruz Biotechnology, 1:2000 dilution), ZSCAN5D (Aviva, 1:2000 dilution) are described in

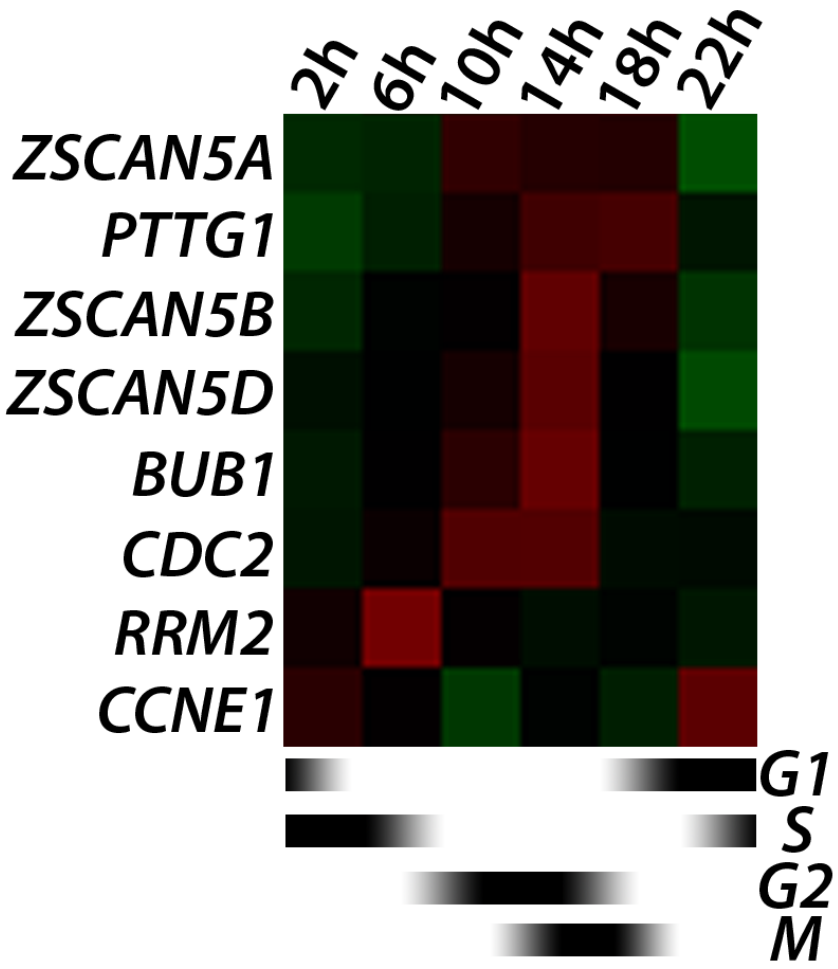
detail and were validated in our previous study (Sun et al., 2016). Anti-Gamma-tubulin (Abcam, 1:5000 dilution), and anti-Phospho-Histone H3 (Ser10) (EMD Millipore, 1:5000 dilution) were purchased from the above named companies. Secondary antibodies targeting primary antibodies were either conjugated with Alexa Fluor® 568 (1:5000 dilution) or Alexa Fluor® 488 (1:5000 dilution) and purchased from Thermo Scientific. Mounting was done using ProLong® Gold Antifade Mountant with DAPI (Thermo Scientific) and confocal microscopy imaging was done using confocal microscope Zeiss LSM880 in the Institute for Genomic Biology Core Facility at the University of Illinois at Urbana-Champaign. More than 10 random images from each sample were processed and counted of total and mitotic cells using ImageJ software (Schneider et al., 2012).

FLOW CYTOMETRY

In order to measure the effects of ZSCAN5 on G1/S/G2 distribution, flow cytometry was done using above described transgenic cell lines stained with propidium iodide. Cells were grown as described above with 1 µg/mL Dox for 48 hours. Each sample was trypsinized using 0.25% trypsine/EDTA (Gibco) and washed with 1X PBS with 0.2% FBS. Cell pellets were fixed with cold ethanol and stored at -20ºC for 24 hours. Fixed cells were washed twice with 1X PBS and stained with FxCycle™ PI/RNase Staining Solution (Thermo Scientific) at room temperature for 30 minutes. BD LSRII Flow cytometry analyzer was used at the Flow Cytometry facility at the University of Illinois at Urbana-Champaign and resulting data were processed and analyzed using FCS Express 5 software.

FIGURES

Figure 3.1.



	2h	6h	10h	14h	18h	22h
ZSCAN5A	475.65	490.855	723.315	695.06	690.885	374.59
ZSCAN5B	25.595	34.48	36.125	59.69	41.675	22.26
ZSCAN5C	8.875	18.62	12	18.025	22.675	9.21
ZSCAN5D	99.605	106.95	118.43	151.965	108.4	71.235
CCNE1	6762.175	5984.81	4737.2	5808.575	5270.755	7847.29
RRM2	39861.63	64851.08	36734.89	32082.56	34015.94	29267.47
CDC2	5326.05	7918.475	13412.69	13568.86	6124.265	6173.055
BUB1	16037.34	25008.09	38027.1	57032.77	23263.43	13312.39
PTTG1	10868.12	15055.15	23568.57	30041.94	30999.69	16495.17

Figure 3.1. (cont.) Cell cycle stage-specific expression ZSCAN5 family genes
Heat map was created based on the expression value (table) acquired from qRT-PCR of time-series sample collections of HEK-293 cells after the release from double thymidine blocks (see methods). Expression levels of *CCNE1*, *RRM2*, *CDC2*, *BUB1* and *PTTG1* were measured on the same cDNA samples to monitor progression through the cell cycle.

Figure 3.2.

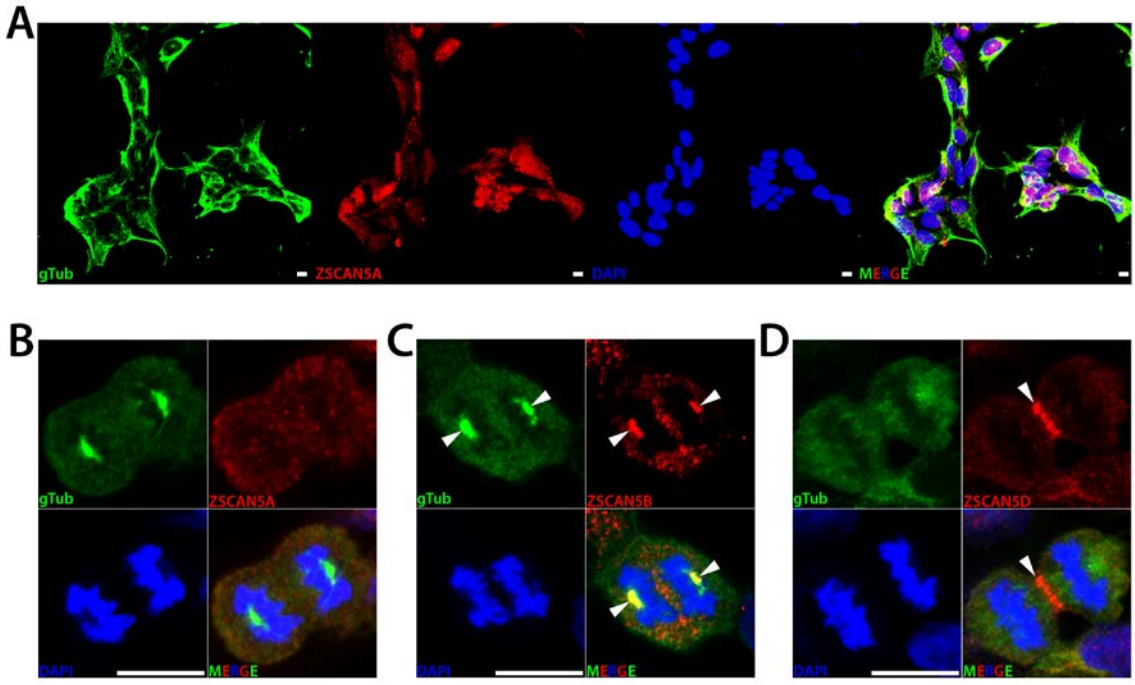

A
gTub
ZSCAN5A
DAPI
MERGE
B
gTub
ZSCAN5A
DAPI
MERGE
C
gTub
ZSCAN5B
DAPI
MERGE
D
gTub
ZSCAN5D
DAPI
MERGE

Figure 3.2. (cont.) Differential cellular localization of ZSCAN5 during cell cycle
HEK-293 cells were grown on glass cover slips and fixed, stained with anti-gamma-tubulin and respective anti-ZSCAN5 antibodies. (A) Selective localization of ZSCAN5A (Alexa Fluor® 568: red) in dividing cells. (B) Dispersed localization of ZSCAN5A (Alexa Fluor® 568: red) during anaphase. (C) Co-localization of ZSCAN5B (Alexa Fluor® 568: red) and centrosome (Alexa Fluor® 488: green, arrowheads) during anaphase. (D) Concentration of ZSCAN5D (Alexa Fluor® 568: red) in the cleavage furrow (arrowhead) in late anaphase. White scale bar corresponds to 10 µm.

Figure 3.3.

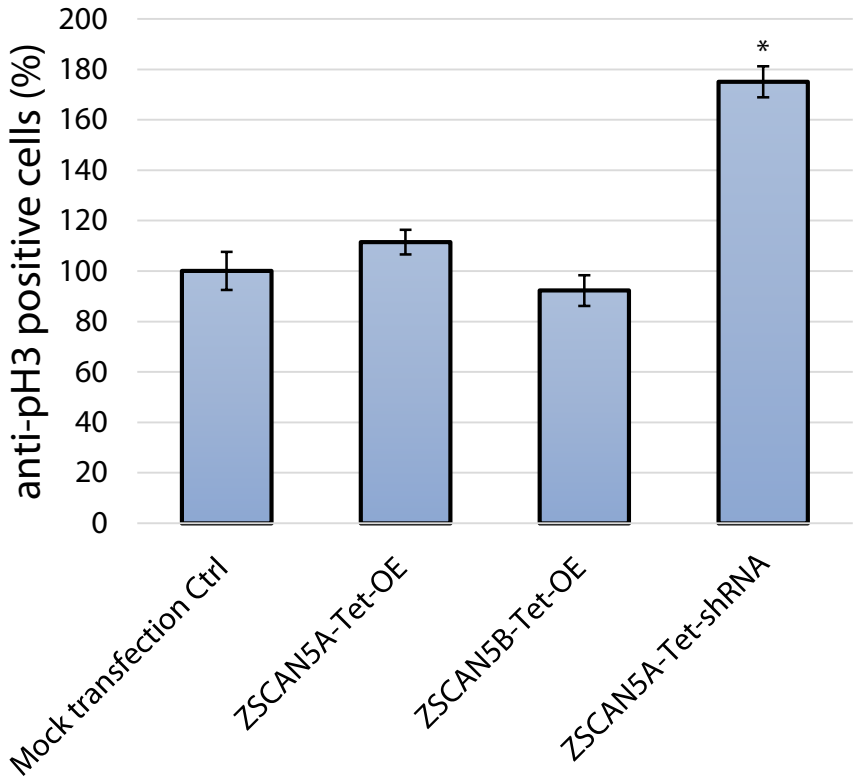

anti-pH3 positive cells (%)
200
180
160
140
120
100
80
60
40
20
0
*
Mock transfection Ctrl
ZSCAN5A-Tet-OE
ZSCAN5B-Tet-OE
ZSCAN5A-Tet-shRNA

Figure 3.3. (cont.) ZSCAN5A knockdown increases numbers of M-phase cells Stably transfected doxycycline-inducible HEK-293 cell lines overexpressing full-length ZSCAN5A (ZSCAN5A-Tet-OE), ZSCAN5B (ZSCAN5B-Tet-OE) or expressing short hairpin RNA targeting ZSCAN5A (ZSCAN5A-Tet-shRNA), or mock transfection control were incubated with 1 µg/mL doxycycline for 48 hours on gelatin-coated glass cover slips. Fixed coverslips were stained with mitotic marker anti-Phospho-Histone H3 (Ser10) antibody, and DAPI to visualize mitotic cells using microscopy. Resulting image was processed and counted of total nuclei and mitotic nuclei using ImageJ software (REF). The proportion of anti-Phospho histone 3 (anti-pH3) positive nuclei or mitotic nuclei out of counted total nuclei after doxycycline induction (Dox+) in mock transfection control was set to 100%. Error bars indicate standard errors for each sample ($SE = \sqrt{\left[\frac{p(1-p)}{n}\right]}$, where *p*, proportion of mitotic nuclei, *n*, total number of nuclei stained). Asterisk refers to statistical significance.

Figure 3.4.

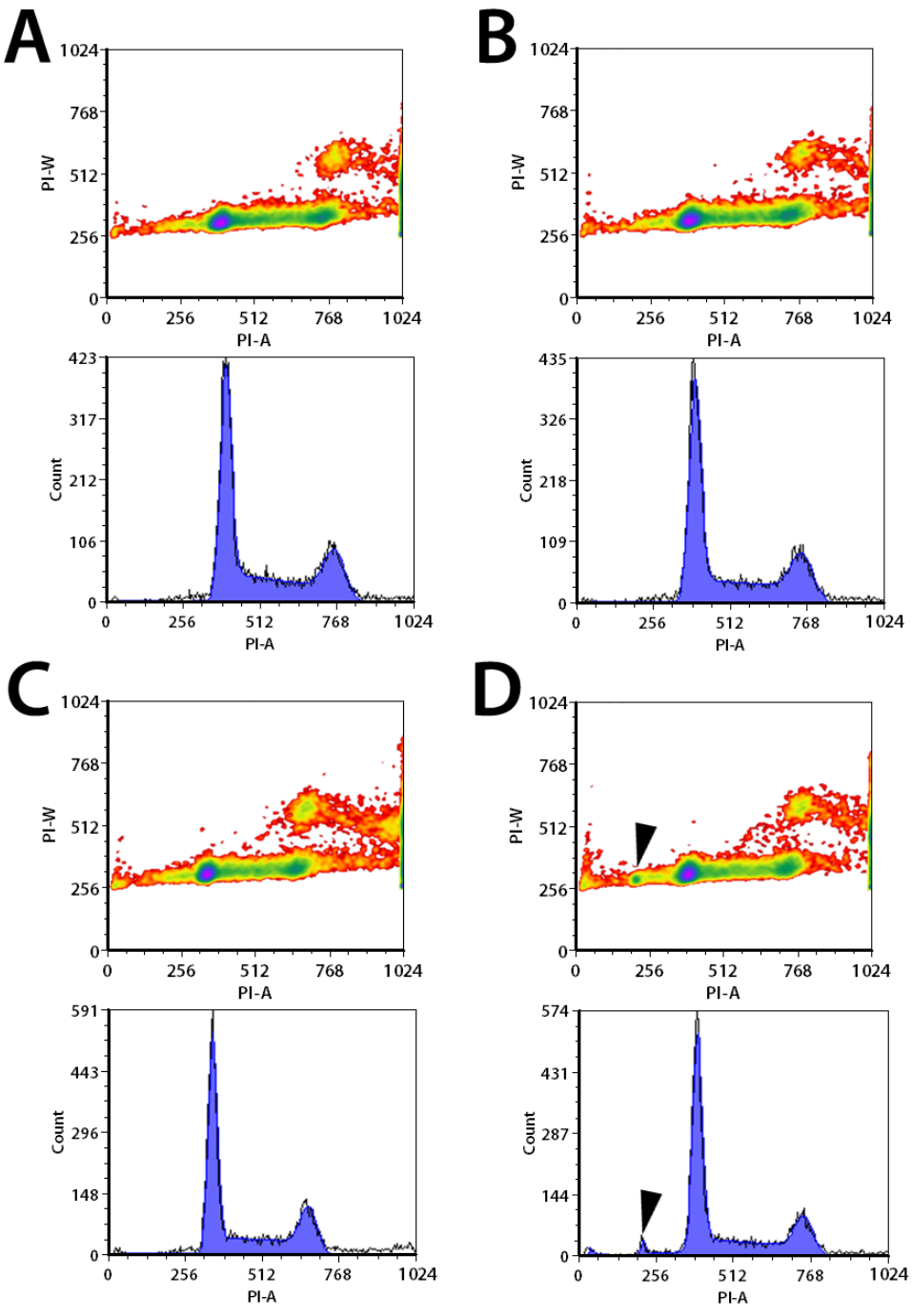

A
B
C
D
PI-W
PI-A
Count
0
256
512
768
1024
106
212
317
423
109
218
326
435
144
287
431
574
148
296
443
591

Figure 3.4. (cont.) Knockdown of ZSCAN5A result in small subset of cells with low DNA content

Flow cytometry was done on inducible transgenic HEK-293 cells stably transfected with (A) mock plasmid control, (B) ZSCAN5A overexpression plasmid (ZSCAN5A-Tet-OE), (C) ZSCAN5B overexpression plasmid (ZSCAN5B-Tet-OE), (D) ZSCAN5A shRNA plasmid (ZSCAN5A-Tet-shRNA) were cultured for 48 hours with addition of 1µg/mL doxycycline. Black arrowheads indicate a small population of cells with lower DNA content compared to normal G1 cells.

TABLE

Table 3.1. Primers and oligonucleotides used in this study

Gene expression primers	DNA sequence (5'>>3')	Amplicon size	Description
ZSCAN5A-F	ACACAAGAGTATTGACGTAACAGGTGA	498	human ZSCAN5A (Tm: 60.6ºC)
ZSCAN5A-R	GCTTGGAATTACACGTAAACCTCTTCTCG		human ZSCAN5A (Tm: 62.2ºC)
ZSCAN5B-F	AAGAGTCCCACAGATCTGGTGAG	388	human ZSCAN5B (Tm: 59.7ºC)
ZSCAN5B-R	TGCTTAGCTGGGAAAAATACTTAAATGA TTTATTG		human ZSCAN5B (Tm: 61.1ºC)
ZSCAN5C-F	TTCCAAACAGTCCCACAGGGG	468	human ZSCAN5C (Tm: 60.6ºC)
ZSCAN5C-R	CATGCAGCGCATAGGCCTG		human ZSCAN5C (Tm: 59.6ºC)
ZSCAN5D-F	CTGTGGTCAATTTTCTTGGCAAGGA	367	human ZSCAN5D (Tm: 60.4ºC)
ZSCAN5D-R	CCCACCACACCTGTAGGACC		human ZSCAN5D (Tm: 60.1ºC)
hYWHAZ-F	ACTTTTGGTACATTGTGGCTTCAA	95	human internal reference control
hYWHAZ-R	CCGCCAGGACAAACCAGTAT		human YWHAZ (Tm: 58ºC)

Cloning primers	DNA sequence (5'>>3')	Tm	
ZSCAN5A-Tet-OE-F	NNNNN-GGATCC-GCCACC ATGGCTGCAAATTGCACATCCTC	60	5'overhang-*Bam*HI-Kozak-DNA sequence
ZSCAN5A-Tet-OE-R	NNNNN GCGGCCGC CTGAGAAGTAGCTTCTGGATG TGTTTTC	60	5'overhang-*Not*I-DNA sequence
ZSCAN5B-Tet-OE-F	NNNNN-GAATTC-GCCACC ATGGCTGCAAATTGCACATCCTC	60	5'overhang-*Eco*RI-Kozak- DNA sequence
ZSCAN5B-Tet-OE-R	NNNNN GCGGCCGC CTGGGAGGTGGTTTCCCG	60	

	ZSCAN5A shRNA oligonucleotide sequences 5' – *Bgl*II – sense – hairpin – antisense – *Xho*I – 3'	**Size**	
ZSCAN5A-Tet-shRNA-sense	GATCCCCGAGAAGAGGTTTACGTGTATT CAAGAGATACACGTAAACCTCTTCTCTTT TTA	60	siRNA Cat. # (Qiagen): SI00779436
ZSCAN5A-Tet-shRNA-antisense	TCGATAAAAAGAGAAGAGGTTTACGTGT ATCTCTTGAATACACGTAAACCTCTTCTC GGG	60	

REFERENCES CITED

1. Bachman, N., Gelbart, M.E., Tsukiyama, T., and Boeke, J.D. (2005). TFIIIB subunit Bdp1p is required for periodic integration of the Ty1 retrotransposon and targeting of Isw2p to S. cerevisiae tDNAs. Genes Dev *19*, 955-964.

2. Barski, A., Chepelev, I., Liko, D., Cuddapah, S., Fleming, A.B., Birch, J., Cui, K., White, R.J., and Zhao, K. (2010). Pol II and its associated epigenetic marks are present at Pol III-transcribed noncoding RNA genes. Nat Struct Mol Biol *17*, 629-634.

3. Bostock, C.J., Prescott, D.M., and Kirkpatrick, J.B. (1971). An evaluation of the double thymidine block for synchronizing mammalian cells at the G1-S border. Exp Cell Res *68*, 163-168.

4. Canella, D., Praz, V., Reina, J.H., Cousin, P., and Hernandez, N. (2010). Defining the RNA polymerase III transcriptome: Genome-wide localization of the RNA polymerase III transcription machinery in human cells. Genome Res *20*, 710-721.

5. Decordier, I., Cundari, E., and Kirsch-Volders, M. (2008). Survival of aneuploid, micronucleated and/or polyploid cells: crosstalk between ploidy control and apoptosis. Mutat Res *651*, 30-39.

6. Dittmar, K.A., Goodenbour, J.M., and Pan, T. (2006). Tissue-specific differences in human transfer RNA expression. PLoS Genet *2*, e221.

7. Donze, D., Adams, C.R., Rine, J., and Kamakaka, R.T. (1999). The boundaries of the silenced HMR domain in Saccharomyces cerevisiae. Genes & development *13*, 698-708.

8. Donze, D., and Kamakaka, R.T. (2001). RNA polymerase III and RNA polymerase II promoter complexes are heterochromatin barriers in Saccharomyces cerevisiae. The EMBO journal *20*, 520-531.

9. Ebersole, T., Kim, J.-H., Samoshkin, A., Kouprina, N., Pavlicek, A., White, R.J., and Larionov, V. (2011). tRNA genes protect a reporter gene from epigenetic silencing in mouse cells. Cell Cycle *10*, 2779-2791.

10. Eisenberg, E., and Levanon, E.Y. (2003). Human housekeeping genes are compact. Trends in genetics : TIG *19*, 362-365.

11. Free, R.B., Hazelwood, L.A., and Sibley, D.R. (2009). Identifying novel protein-protein interactions using co-immunoprecipitation and mass spectroscopy. Curr Protoc Neurosci *Chapter 5*, Unit 5 28.

12. Gaither, T.L., Merrett, S.L., Pun, M.J., and Scott, K.C. (2014). Centromeric barrier disruption leads to mitotic defects in Schizosaccharomyces pombe. G3 (Bethesda) *4*, 633-642.

13. Gelbart, M.E., Bachman, N., Delrow, J., Boeke, J.D., and Tsukiyama, T. (2005). Genome-wide identification of Isw2 chromatin-remodeling targets by localization of a catalytically inactive mutant. Genes & development *19*, 942-954.

14. Haeusler, R.A., Pratt-Hyatt, M., Good, P.D., Gipson, T.A., and Engelke, D.R. (2008). Clustering of yeast tRNA genes is mediated by specific association of condensin with tRNA gene transcription complexes. Genes & development *22*, 2204-2214.

15. Hakimi, M.-A., Bochar, D.A., Schmiesing, J.A., Dong, Y., Barak, O.G., Speicher, D.W., Yokomori, K., and Shiekhattar, R. (2002). A chromatin remodelling complex that loads cohesin onto human chromosomes. Nature *418*, 994-998.

16. Harismendy, O., Gendrel, C.G., Soularue, P., Gidrol, X., Sentenac, A., Werner, M., and Lefebvre, O. (2003). Genome-wide location of yeast RNA polymerase III transcription machinery. The EMBO Journal *22*, 4738-4747.

17. Huang, Y., and Maraia, R.J. (2001). Comparison of the RNA polymerase III transcription machinery in Schizosaccharomyces pombe, Saccharomyces cerevisiae and human. Nucleic Acids Research *29*, 2675-2690.

18. Hull, M.W., Erickson, J., Johnston, M., and Engelke, D.R. (1994). tRNA genes as transcriptional repressor elements. Molecular and cellular biology *14*, 1266-1277.

19. Kendall, A., Hull, M.W., Bertrand, E., Good, P.D., Singer, R.H., and Engelke, D.R. (2000). A CBF5 mutation that disrupts nucleolar localization of early tRNA biosynthesis in yeast also suppresses tRNA gene-mediated transcriptional silencing. Proceedings of the National Academy of Sciences *97*, 13108-13113.

20. Kinsey, P.T., and Sandmeyer, S.B. (1991). Adjacent pol II and pol III promoters: transcription of the yeast retrotransposon Ty3 and a target tRNA gene. Nucleic acids research *19*, 1317-1324.

21. Marshall, L., and White, R.J. (2008). Non-coding RNA production by RNA polymerase III is implicated in cancer. Nat Rev Cancer *8*, 911-914.

22. Mertens, C., and Roeder, R.G. (2008). Different functional modes of p300 in activation of RNA polymerase III transcription from chromatin templates. Mol Cell Biol *28*, 5764-5776.

23. Moqtaderi, Z., and Struhl, K. (2004). Genome-wide occupancy profile of the RNA polymerase III machinery in Saccharomyces cerevisiae reveals loci with incomplete transcription complexes. Molecular and cellular biology *24*, 4118-4127.

24. Moqtaderi, Z., Wang, J., Raha, D., White, R.J., Snyder, M., Weng, Z., and Struhl, K. (2010). Genomic binding profiles of functionally distinct RNA polymerase III transcription complexes in human cells. Nature structural & molecular biology *17*, 635-640.

25. Musacchio, A. (2015). The Molecular Biology of Spindle Assembly Checkpoint Signaling Dynamics. Curr Biol *25*, R1002-1018.

26. Oler, A.J., Alla, R.K., Roberts, D.N., Wong, A., Hollenhorst, P.C., Chandler, K.J., Cassiday, P.A., Nelson, C.A., Hagedorn, C.H., and Graves, B.J. (2010). Human RNA polymerase III transcriptomes and relationships to Pol II promoter chromatin and enhancer-binding factors. Nature structural & molecular biology *17*, 620-628.

27. Parelho, V., Hadjur, S., Spivakov, M., Leleu, M., Sauer, S., Gregson, H.C., Jarmuz, A., Canzonetta, C., Webster, Z., and Nesterova, T. (2008). Cohesins functionally associate with CTCF on mammalian chromosome arms. Cell *132*, 422-433.

28. Raab, J.R., Chiu, J., Zhu, J., Katzman, S., Kurukuti, S., Wade, P.A., Haussler, D., and Kamakaka, R.T. (2012). Human tRNA genes function as chromatin insulators. The EMBO journal *31*, 330-350.

29. Roberts, D.N., Stewart, A.J., Huff, J.T., and Cairns, B.R. (2003). The RNA polymerase III transcriptome revealed by genome-wide localization and activity–occupancy relationships. Proceedings of the National Academy of Sciences *100*, 14695-14700.

30. Schneider, C.A., Rasband, W.S., and Eliceiri, K.W. (2012). NIH Image to ImageJ: 25 years of image analysis. Nat Methods *9*, 671-675.

31. Snider, C.E., Stephens, A.D., Kirkland, J.G., Hamdani, O., Kamakaka, R.T., and Bloom, K. (2014). Dyskerin, tRNA genes, and condensin tether pericentric chromatin to the spindle axis in mitosis. J Cell Biol *207*, 189-199.

32. Thompson, M., Haeusler, R.A., Good, P.D., and Engelke, D.R. (2003). Nucleolar clustering of dispersed tRNA genes. Science *302*, 1399-1401.

33. Wang, L., Haeusler, R.A., Good, P.D., Thompson, M., Nagar, S., and Engelke, D.R. (2005). Silencing near tRNA genes requires nucleolar localization. Journal of Biological Chemistry *280*, 8637-8639.

34. White, R.J., Gottlieb, T.M., Downes, C.S., and Jackson, S.P. (1995). Cell cycle regulation of RNA polymerase III transcription. Mol Cell Biol *15*, 6653-6662.

35. Whitfield, M.L., Sherlock, G., Saldanha, A.J., Murray, J.I., Ball, C.A., Alexander, K.E., Matese, J.C., Perou, C.M., Hurt, M.M., Brown, P.O., *et al.* (2002). Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Molecular biology of the cell *13*, 1977-2000.

36. Wuelling, M., Pasdziernik, M., Moll, C.N., Thiesen, A.M., Schneider, S., Johannes, C., and Vortkamp, A. (2013). The multi zinc-finger protein Trps1 acts as a regulator of histone deacetylation during mitosis. Cell Cycle *12*, 2219-2232.

CHAPTER 4:

TOOLS AND RESOURCES FOR ANALYSIS OF BIOLOGICAL FUNCTIONS OF MOUSE *Zscan5b*

Younguk Sun[1,2], and Lisa Stubbs[1,2,3]

ABSTRACT

The results presented in this thesis suggest an important role for the conserved gene *ZSCAN5B* and its primate-specific paralogs in the regulation of transcription of RNA Polymerase III (Pol III) loci as well as neighboring Pol II genes. Gene knockdown and genetic manipulation experiments have indicated that human *ZSCAN5A* plays an essential role in mitotic progression and chromosome integrity, and suggested similar and cooperative functions for *ZSCAN5B*. Gene knockdown experiments and *in vivo* expression data also suggested that the ZSCAN5 genes play a role in differentiation decisions in a number of tissues. Together these data suggest a number of intriguing hypotheses for future studies that will require the development of specialized resources.

In this chapter, I will describe the generation of some essential tools, including expression constructs and two sets of transgenic mouse lines that I have already developed for this purpose. I also outline ideas for experiments, some of which are already underway, that can use these resources to address the root biological functions of the ZSCAN5 family through analysis of the unique mouse family member, *Zscan5b*.

[1] Institute for Genomic Biology

[2] Department of Cell and Developmental Biology,
University of Illinois at Urbana-Champaign, Urbana IL 61801

[3] Corresponding author

INTRODUCTION

This thesis has focused on the unique eutherian gene, *ZSCAN5B*, which duplicated to give rise to clustered paralogs, *ZSCAN5A, ZSCAN5C and ZSCAN5D*, in early primate history. In a previous chapter (Chapter 2), I showed that three of the primate genes, ZSCAN5A, B, and D, are expressed in different human tissues and are robustly expressed in rapidly dividing cell populations. I also showed that the ZSCAN5 proteins bind preferentially to RNA Polymerase III (Pol III) transcription units, especially tDNAs, and related repetitive elements. ZSCAN5B protein binds with high specificity to a particular subset of tDNA sequences in two human cell lines, and to the same subset of conserved tDNA in mouse fetal placenta. Furthermore, ZSCAN5A and ZSCAN5D also occupy RNA Polymerase III (Pol III) promoters as well as extra-TFIIIC (ETC) sites throughout the genome.

Combining ChIP-seq data with the analysis of global gene expression after ZSCAN5 gene knockdown, I showed that ZSCAN5 protein binding not only regulates expression of the bound Pol III genes, but also affects the expression of neighboring Pol II genes implicated in ribosome biogenesis, tRNA processing, cell cycle progression and cellular differentiation. In particular, gene expression data implicated ZSCAN5A for a role in chromosome segregation and in Chapter 3, I further investigated this hypothesis by manipulating the levels of *ZSCAN5A* expression in mammalian cells. Consistent with an essential role in cell cycle progression, *ZSCAN5A* knockdown led to the accumulation of cells in the mitotic M-phase and the appearance of aneuploid cells. Since gene expression and ChIP data also suggested that *ZSCAN5B* cooperates with *ZSCAN5A* in regulating cell cycle regulation and differentiation (Chapter 2), it follows that the conserved mouse gene might play a similar role.

In this chapter, I describe the development tools to further investigate the biological roles of the ancestral *Zscan5b* in the mouse model system, including inducible overexpression and shRNA plasmid constructs. These constructs have been used to establish stable cell lines and importantly, transgenic mouse lines that will allow future investigations of the developmental roles of *Zscan5b* and the effects of gene knockdown or overexpression on mouse health, fertility and behavior.

RESULTS

TOOLS FOR FURTHUR STUDY

Inducible overexpression of mouse Zscan5b, mZscan5b-Tet-OE

My first goal was to generate constructs to allow the expression of the Zscan5b protein in mouse cells and transgenic mice. From data reported in Chapters 2 and 3, I reasoned that these expression constructs should be inducible rather than constitutively expressed, in order to prevent any possible negative effects on cell maintenance and transgenic animal viability. I therefore chose the tetracycline-inducible (Tet-on) system to generate these expression constructs (Gossen et al., 1995).

The pTRE-Tight vector is designed with a tetracycline response element (TRE) promoter that is tightly regulated but highly inducible in the presence of Doxycycline (Dox), and a SV40 poly-A signal for a stable translation. I further modified this plasmid to incorporate two STREP-tag II (WSHPQFEK) (Korndorfer and Skerra, 2002) and three hemagglutinin tags (HA: YPYDVPDYA) (Wilson et al., 1984) to the C-terminus of the protein of interest and named "pTTS2.1". The tags are incorporated into the expressed protein to provide a tool for efficient detection and purification, as described in more detail the Discussion.

This pTRE-Tight plasmid can overexpress the cloned gene only in the presence of reverse Tet Trans-Activator (rtTA) protein and Dox (Gossen et al., 1995). In order to investigate the biological functions of the mouse version of *Zscan5b* (*mZscan5b* hereafter) as they relate to cellular functions and differentiation, I created a Tet-inducible plasmid construct capable of overexpressing the open reading frame of full-length *mZscan5b*. After thorough validation of its DNA sequences, the resulting plasmid was named "mZscan5b-Tet-OE", purified and stored for the downstream experiments such mammalian cell transfections and transgenic mouse line establishment.

Inducible small hairpin RNA (shRNA) targeting mZscan5b, mZscan5b-Tet-shRNA

In contrast to the inducible overexpression plasmid construct described above, I also created a version of inducible plasmid that conditionally expresses shRNA upon Dox induction in order to assess the phenotypic difference compared to mZscan5b-Tet-OE. As effective the gene ablation using siRNA treatment as it may be, the treatment has its drawback due to the instability of the siRNA and thus short duration of knockdown, making it unsuited to evaluate the long-term consequences of transcript depletion. I used an inducible shRNA plasmid with embedded mammalian selective marker, pSuperior.Puro (Oligoengine), enabling its stable incorporation into genome for later experiments (Brummelkamp et al., 2002a, b). The insert shRNA sequence was based on the siRNA sequence that was most efficient at mZscan5b knockdown among a set of three siRNA designs (purchased from Qiagen, Inc.; data not shown). This plasmid was named "mZscan5b-Tet-shRNA", and thoroughly sequenced and purified for further experimentation.

Transfection with mZscan5b-Tet-OE and mZscan5b-Tet-shRNA to mouse cell lines

Cloned and purified plasmids were initially tested by transfection to the Neuro2a Tet-on cell line, a modified mouse neuroblastoma cell line that constitutively expresses rtTA (Clontech). Transfected cells were selected under puromycin for 14 days and individual colonies were subcultured to establish stably transfected cell lines. Each colony was genotyped and analyzed for transgene induction efficiency, or for the efficiency of mZscan5b knockdown after Doxycycline (Dox) treatment compared to uninduced cells or vector-only transfected controls. Transcript level analysis was done using qRT-PCR, and mZscan5b protein expression was measured using Western blots. This transfection analysis was used as an intermediary testing platform to validate the Tet-inducibility for the next step, the establishment of transgenic mice.

Generation of transgenic mice overexpressing and knocking down mZscan5b

The investigation of biological functions of *mZscan5b in vivo* is our paramount interest. To enable this investigation, the functional fragments of the plasmid including the Tet-regulatable promoter, the Strep-and HA-tagged mZscan5b open reading frame,

and the SV40 poly-A signal DNA sequence from the mZscan5b-Tet-OE plasmid was cleaved, gel isolated and employed to create transgenic mouse lines that can conditionally overexpress *mZscan5b*.

At the same time, the mZscan5b-Tet-shRNA plasmid was also linearized to make another set of transgenic mouse lines able to express the shRNA targeting *mZscan5b* upon Dox induction. Both constructs were injected into wild type mice oocytes at the Transgenic Mouse Facility at University of Illinois at Urbana-Champaign. Initial screening was done by tail-snip genotyping of pups from a number of transgenic founders of both sets of lines: *mZscan5b* overexpression mice (mZ5b-FL) and *mZscan5b* shRNA mice (mZ5b-sh).

For mZ5b-FL, only 1 out of 37 male mice and 8 out of 32 female mice were positive for transgene incorporation, while 4 out of 27 male mice and 5 out of 17 female mice showed strong positive signals for mZ5b-sh. Each positively selected male mouse was then bred with wildtype females to yield offspring and to establish transgenic lines from several founders. These lines are now ready for testing, as described below.

DISCUSSION

As described, I have generated tetracycline inducible over-expression and shRNA constructs to study the functions of *Zscan5b* in the mouse model system. These inducible protein constructs offer several advantages that we can now employ in studies of mZscan5b function. First, we can culture cells or breed mice without expressing the transgene, then induce their expression by adding Dox to the culture medium or feeding the animals Dox. In Tet-on transgenes, protein or shRNA expression is proportional to the concentration of Dox administered (Seibler et al., 2007). This fact enable us to “fine tune” the dosage of mZscan5b in the cell lines or mice.

Second, our constructs are designed to attach widely used peptide tags to the proteins, Strep-tag II and HA, that can be detected with commercially available, high-specificity antibodies; this creates an opportunity for follow-on experiments of several types. For example, we are interested in identifying cellular binding partners for Zscan5b, and these tags will allow for very efficient methods of protein isolation, and for more

complex protocols like co-immunoprecipitation followed by mass spectroscopy (Free et al., 2009). We were able to generate an antibody targeting mZscan5b and use it for Chromatin immunoprecipitation (Chapter 2), but this antibody did not show the type of high affinity we can expect from anti-HA or anti-FLAG reagents.

Furthermore, purification of C2H2 zinc finger proteins in *E.* coli has been a huge challenge, due to the nature of cysteine-cysteine disulfide bridges which leads to the proteins being retained in inclusion bodies after overexpression (Yan and Sun, 2009). For that reason, expression and purification of transgene in a mammalian cell line, which is already equipped with proper protein folding or re-folding machinery such as molecular chaperones, has been an alternative and preferred solution (Hurt et al., 2003). Purification of mZscan5b proteins after mammalian expression will be facilitated by the Strep-tag II, owing to its high-specificity. The purified protein can then be used to determine the binding affinity or specificity to certain DNA sequences using electrophoretic mobility shift assay (EMSA) or other protein partners *in vitro*.

Most importantly, the transgenic mice I have generate provide new an important resources for the Stubbs laboratory, which can be used in a number of future studies. A few of these experiments are already underway. For example, female transgenic mice have been bred to male animals carrying a constitutive rtTA gene inserted into the ROSA-26 locus (*Gt(ROSA)26Sor*$^{tm1(rtTA*M2)Jae}$/J; from The Jackson Laboratory) (Hochedlinger et al., 2005). If a transgenic female is crossed to a male ROSA-rtTA homozygous male and fed Dox during pregnancy, the *mZscan5b* transgenes will be activated in 50% of her pups (the pups inheriting both mZscan5b and rtTA transgenes). These experiments will allow us to assess the role of *Zscan5b* on mouse development. Since the Dox can be administered at different times during pregnancy, we can also target certain stages of development.

Another set of experiments, now underway, will test the role of *mZscan5b* in mouse fertility. As described in Chapter 2, *mZscan5b* is very highly expressed in testis and IHC studies have shown that the protein is present in developing germ cells. To test the role in germ cell differentiation, we are treating males with the mZscan5b and ROSA-rtTA transgenes with Dox postnatally and examining the progression of spermatogenesis.

These resources thus allow us to investigate the biological functions of mouse *Zscan5b* both during development and in adult tissues. These new studies will permit us to place the regulation of tDNA, Pol III transcripts, and neighboring genes in the context of development and tissue maintenance *in vivo.*

MATERIALS AND METHODS

PLASMIDS AND CLONING

Modification of Tet-inducible overexpression plasmid, pTTS2.1

The pTRE-Tight vector was purchased from Clontech (Cat. 631059). For added functionality, two STREP-tag II (WSHPQFEK) and three hemagglutinin (HA: YPYDVPDYA) tags were incorporated to the 3' of multiple cloning sites. To do this, 150-mer of sense and antisense oligonucleotides which encode aforementioned peptide tags flanked by restriction enzyme sites, *Xba*I, were ordered from Integrative DNA Technology (*Table 4.1*). 2 µg of each oligonucleotide was mixed in the annealing buffer (10 mM Tris-HCl pH8.0, 50 mM NaCl, 1 mM EDTA) and heated up to 95^{o}C for 10 mins in a heat block. The heated block was left out at room temperature to slowly cool down to anneal two oligonucleotides. The annealed double stranded DNA was then ligated to *Xba*I-digested pTRE-Tight plasmid. Cloned plasmid was sequenced to verify correct insertion and purified for later experiments. The resulting plasmid was named as "pTTS2.1" plasmid. All primers and oligonucleotides used in this chapter are described in *Table 4.1*. Restriction enzymes and T4 DNA ligase were purchased from New England Biolabs and Pfx High-fidelity PCR DNA polymerase was purchased from Life Technology. PCR purification kit and Plasmid Mini-/Midi- preparation kits were purchased from Qiagen.

Construction of Tet-inducible overexpression plasmid, mZscan5b-Tet-OE

In order to clone the open reading frame (ORF) of *mZscan5b*, testis cDNA from C57BL6/J mouse was obtained and used as a PCR template to generate *mZscan5b* ORF insert DNA flanked by *Eco*RI (5') and *Not*I (3') restriction sites using Pfx High-fidelity

DNA polymerase (Life Technology) and primers (*Table 4.1*). The resulting PCR amplicon was purified, digested with *Eco*RI and *Not*I, and ligated to double-digested pTTS2.1 plasmid. After the validation the DNA sequences, plasmids were purified from *E*. coli using a Midi-prep Plasmid Purification system (Qiagen Cat. 12145), then labeled as “mZscan5b-Tet-OE”.

Construction of Tet-inducible shRNA plasmid, mZscan5b-Tet-shRNA

To create an inducible cell line that expresses short hairpin RNAs (shRNAs) targeting *mZscan5b*, an annealed double stranded oligonucleotides that encode shRNA sequence was cloned into the pSuperior.Puro plasmid (Oligoengine) and named as “mZscan5b-Tet-shRNA”

CELL CULTURE AND DNA TRANSFECTION

Neuro2a Tet-on (Clontech) cells were cultured in Dulbecco's Modified Eagle's Medium (DMEM) containing 2 mM L-glutamine, 10% Tetracycline-free fetal bovine serum (FBS, Clontech), 1X Pen Strep on 0.2% gelatin-coated flasks and incubated at 37 °C in 5% CO2. For plasmid DNA transfection, about $4.5x10^5$ Neuro2a Tet-on cells were seeded to 6-well plates 24 hours before, and 3 µg of each plasmid; mZscan5b-Tet-OE, or blank plasmid along with 300 ng of pPur plasmid (Clontech) which confers puromycin resistance, or mZscan5b-Tet-shRNA was transfected using Lipofectamine 2000 (Invitrogen). 24 hours later, transfected cells were selected under 1 µg/ml of puromycin for additional 14 days. Single colonies were selected and expanded, then tested for the efficiency of ZSCAN5A by qRT-PCR knockdown 48 h after addition of 1 µg/mL Doxycycline (Dox; Sigma-Aldrich). Each colony was analyzed of it overexpression or knockdown of respective genes and proteins compared to Dox-treated cells carrying the empty vector.

RNA PREPARATION AND QUANTITATIVE RT-PCR

Total RNA was isolated from cell lines and tissues using TRIzol (Invitrogen) followed by 30 minutes of RNase-free DNaseI treatment (New England Biolabs) at 37^{o}C and RNA Clean & Concentrator^{TM}-5 (Zymo Research). 2 µg of total RNA was used to

generate cDNA using Superscript III Reverse Transcriptase (Invitrogen) with random hexamers (Invitrogen) according to manufacturer's instructions.

The resulting cDNAs were analyzed of transcript-specific expression through quantitative reverse-transcript PCR (qRT-PCR) using Power SYBR Green PCR master mix (Applied Biosystems, you are supposed to say where the companies are) with custom-designed primer sets (*Table 3.1*) purchased from Integrated DNA Technology. Relative expression was determined by normalizing the expression of all genes of interest to either human or mouse Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, zeta polypeptide (*YWHAZ*) expression (ΔCt) as described (Eisenberg and Levanon, 2003).

PROTEIN PREPARATION AND WESTERN BLOTS

Nuclear Extracts were prepared with NucBuster™ Protein Extraction Kit (Novagen) and measured by Bradford-based assay (BioRad). The extracts were stored at -80^{o}C and thawed on ice with the addition of protease inhibitor Cocktail (Roche) directly before use. 15 µg of nuclear extracts were run on 10% acrylamide gels and transferred to hydrophobic polyvinylidene difluoride (PVDF) membrane (GE-Amersham, 0.45 µm) using BioRad Semi-dry system, then visualized by exposure to MyECL Imager (Thermo Scientific).

FIGURES

Figure 4.1. DNA sequence and plasmid map of Doxycycline-inducible mZscan5b overexpression construct, mZscan5b-Tet-OE

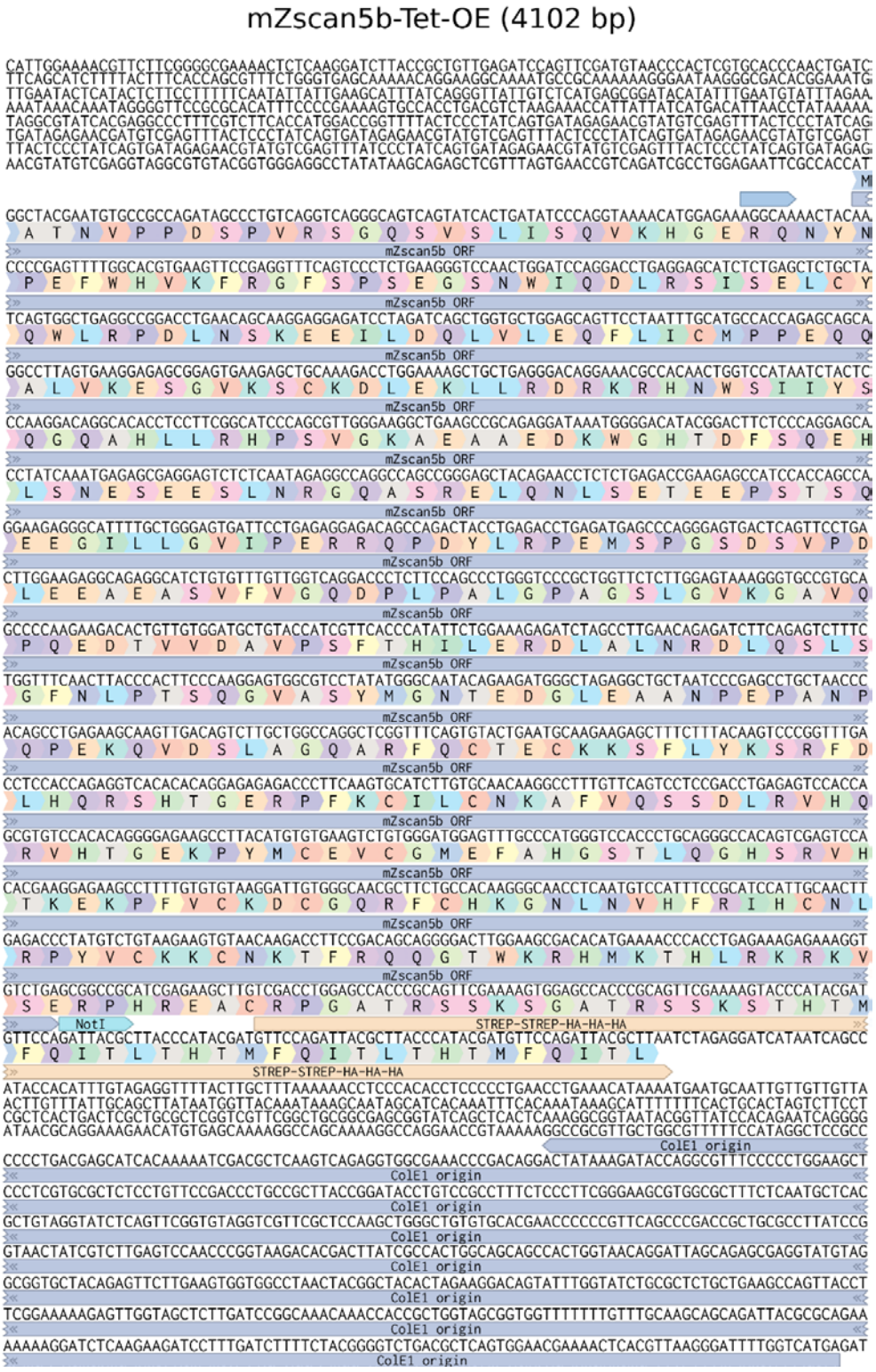


Figure 4.1. (cont.)

TATCAAAAAGGATCTTCACCTAGATCCTTTTAAATTAAAAATGAAGTTTTAAATCAATCTAAAGTATATATGAGTAAACTTGGTCTGACAGTT
ACCAATGCTTAATCAGTGAGGCACCTATCTCAGCGATCTGTCTATTTCGTTCATCCATAGTTGCCTGACTCCCCGTCGTGTAGATAACTACGA
AmpR
TACGGGAGGGCTTACCATCTGGCCCCAGTGCTGCAATGATACCGCGAGACCCACGCTCACCGGCTCCAGATTTATCAGCAATAAACCAGCCAG
AmpR
CCGGAAGGGCCGAGCGCAGAAGTGGTCCTGCAACTTTATCCGCCTCCATCCAGTCTATTAATTGTTGCCGGGAAGCTAGAGTAAGTAGTTCGC
AmpR
CAGTTAATAGTTTGCGCAACGTTGTTGCCATTGCTACAGGCATCGTGGTGTCACGCTCGTCGTTTGGTATGGCTTCATTCAGCTCCGGTTCCC
AmpR
AACGATCAAGGCGAGTTACATGATCCCCCATGTTGTGCAAAAAAGCGGTTAGCTCCTTCGGTCCTCCGATCGTTGTCAGAAGTAAGTTGGCCG
AmpR
CAGTGTTATCACTCATGGTTATGGCAGCACTGCATAATTCTCTTACTGTCATGCCATCCGTAAGATGCTTTTCTGTGACTGGTGAGTACTCAA
AmpR
CCAAGTCATTCTGAGAATAGTGTATGCGGCGACCGAGTTGCTCTTGCCCGGCGTCAATACGGGATAATACCGCGCCACATAGCAGAACTTTAA
AmpR
AAGTGCTCAT
AmpR

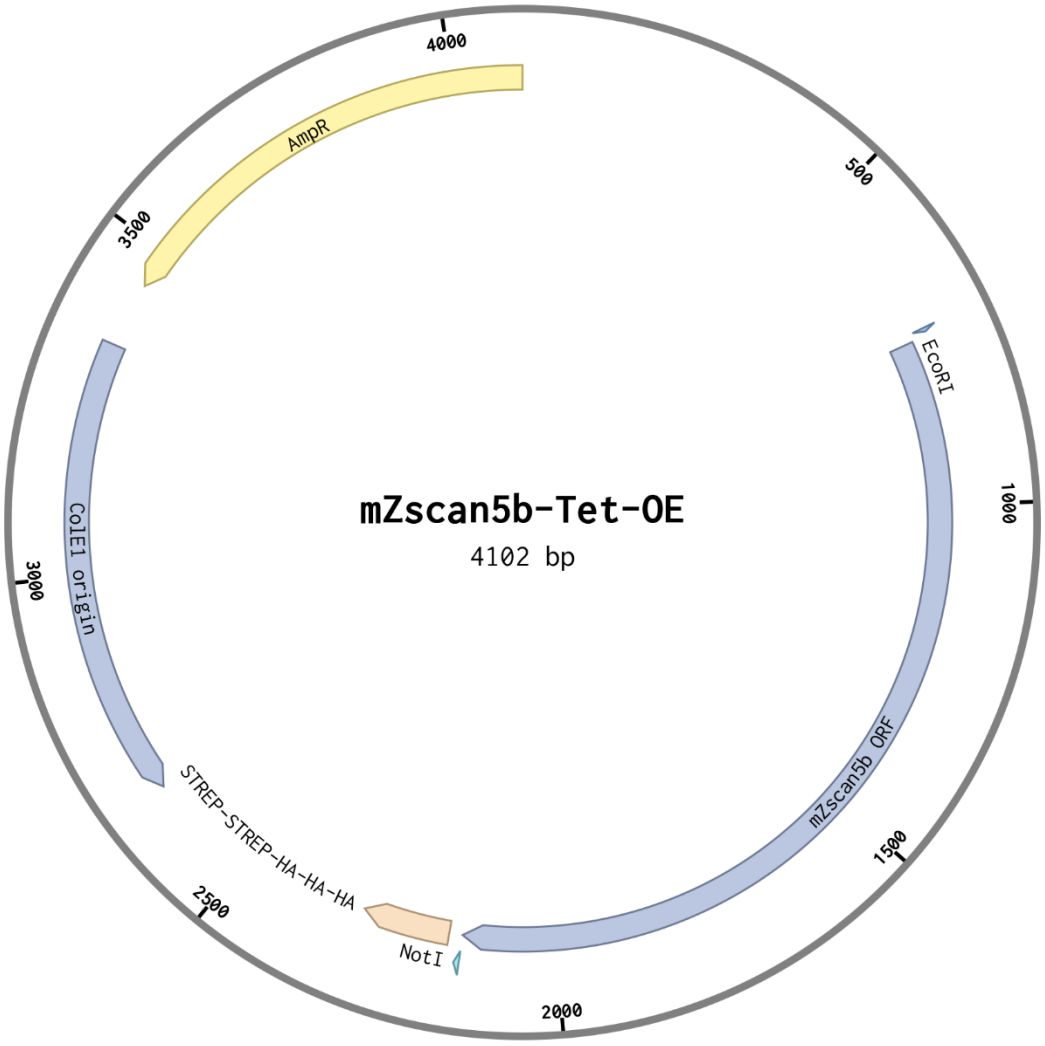


Figure 4.2. DNA sequence and plasmid map of Doxycycline-inducible mZscan5b shRNA construct, mZscan5b-Tet-shRNA

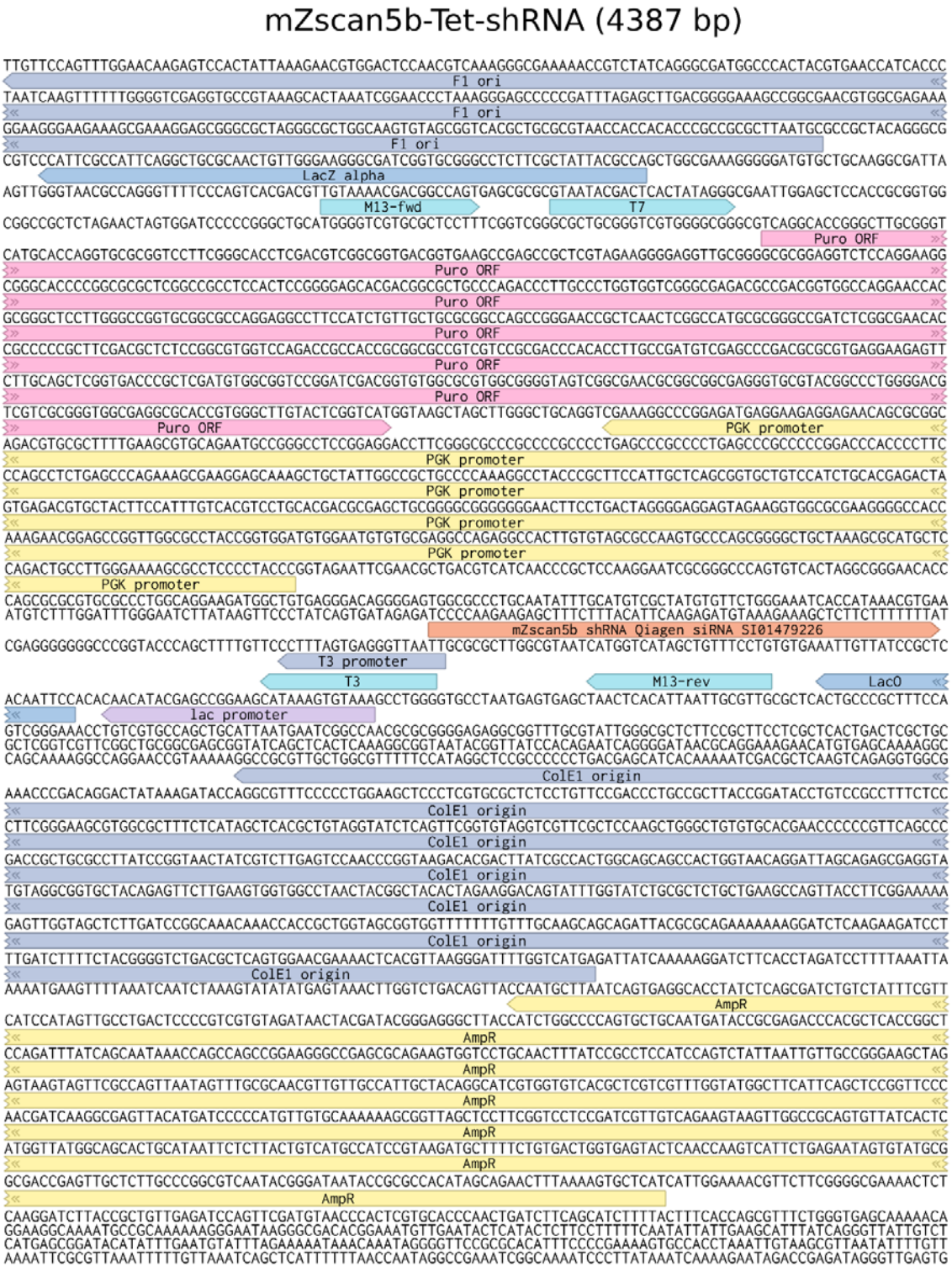
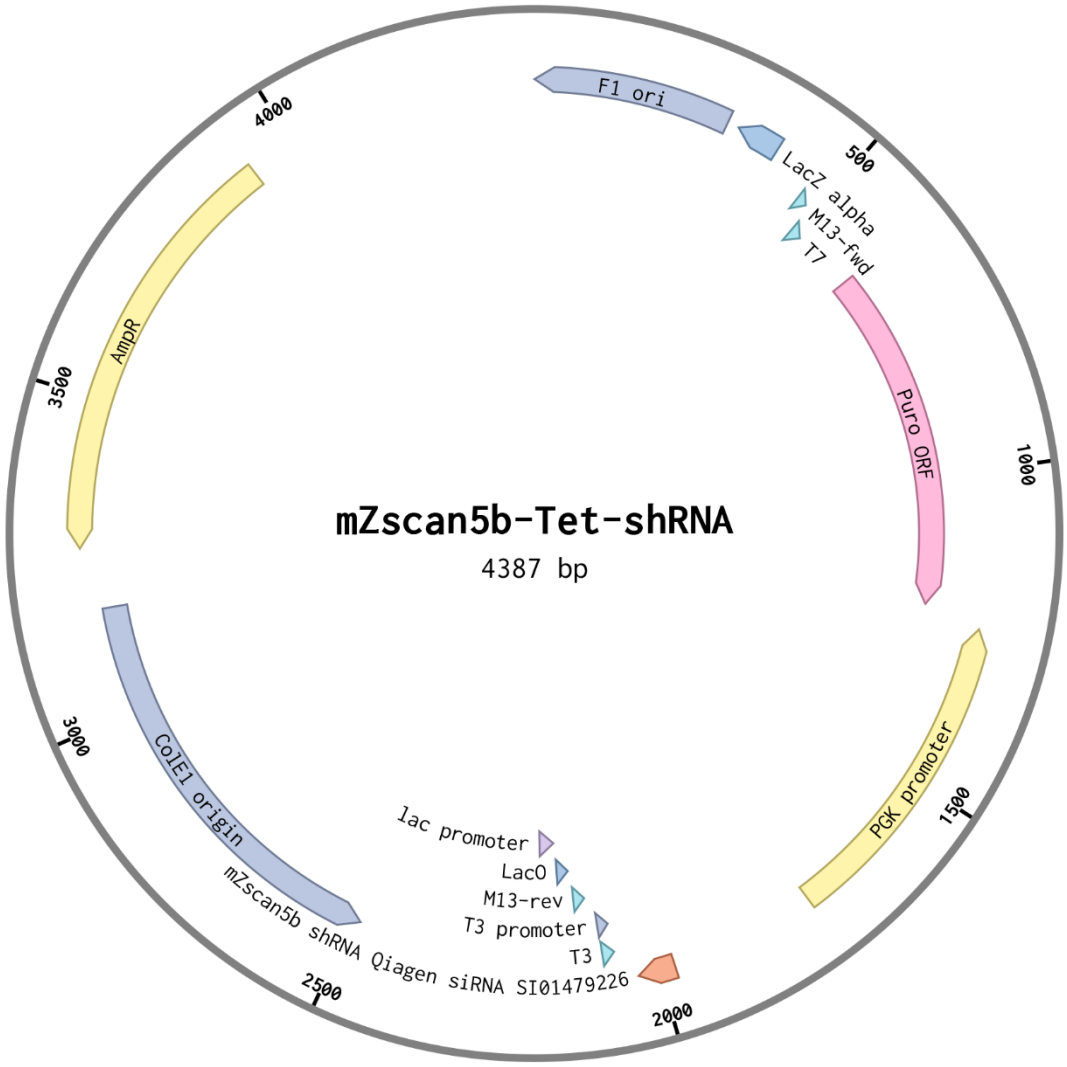


Figure 4.2. (cont.)



TABLE

Table 4.1. Primers and oligonucleotides used in this study

Oligonucleotides	DNA sequence (5'>>3')	Amplicon size	Description
pTTS2.1 Tag sense	CTAGA-C-GGT-GCA- TGGAGCCACCCGCAGTTCGAAAAG- TGGAGCCACCCGCAGTTCGAAAAG- -GGT-GCA- TACCCATACGATGTTCCAGATTACGCT- TACCCATACGATGTTCCAGATTACGCT- TACCCATACGATGTTCCAGATTACGCT- TAA-T	150	*Xba*I-Gly- Ala-Strep- Strep- Gly-Ala- HA-HA- HA-STOP
pTTS2.1 Tag antisense	TCTAGATTAAGCGTAATCTGGAACATCGTA TGGGTAAGCGTAATCTGGAACATCGTATGG GTAAGCGTAATCTGGAACATCGTATGGGTT GCACCCTTTTCGAACTGCGGGTGGCTCCACT TTTCGAACTGCGGGTGGCTCCATGCACCG	150	
	mZscan5b shRNA oligonucleotide sequences 5' – *Bgl*II – sense – hairpin – antisense – *Xho*I – 3'	**Size**	
mZscan5b-Tet-shRNA-sense	GATCCCCAAGAAGAGCTTTCTTTACATTCAA GAGATGTAAAGAAAGCTCTTCTTTTTTA	60	siRNA Cat. # (Qiagen): SI01489226
mZscan5b-Tet-shRNA-antisense	TCGATAAAAA AAGAAGAGCTTTCTTTACA TCTCTTGAATGTAAAGAAAGCTCTTCTTGGG	60	
Cloning primers	**DNA sequence (5'>>3')**	**Tm**	
mZscan5b-Tet-OE-F	NNNNN-GAATTC-GCCACC ATGGCTACGAATGTGCCGCC	61.3	5'overhang-*Eco*RI-Kozak-DNA sequence
mZscan5b-Tet-OE-R	NNNNN GCGGCCGC TCAGACACCTTTCTCTTTCTCAGGTG	60.7	5'overhang-*Not*I-DNA sequence
Genotyping primers	**DNA sequence (5'>>3')**	**Tm**	
mZ5b-FL-F	TGCAGGGCCACAGTCGAG	60	mZ5b-FL genotype
mZ5b-FL-R	CATCGTATGGGTACTTTTCGAACTGC	60.3	mZ5b-FL genotype
mZ5b-sh-F	CATCAACCCGCTCCAAGGAATC	59.6	mZ5b-sh genotype
mZ5b-sh-R	AATGTGAGTTAGCTCACTCATTAGGC	59.4	mZ5b-sh genotype

REFERENCES CITED

1. Brummelkamp, T.R., Bernards, R., and Agami, R. (2002a). Stable suppression of tumorigenicity by virus-mediated RNA interference. Cancer Cell *2*, 243-247.

2. Brummelkamp, T.R., Bernards, R., and Agami, R. (2002b). A system for stable expression of short interfering RNAs in mammalian cells. Science *296*, 550-553.

3. Free, R.B., Hazelwood, L.A., and Sibley, D.R. (2009). Identifying novel protein-protein interactions using co-immunoprecipitation and mass spectroscopy. Curr Protoc Neurosci *Chapter 5*, Unit 5 28.

4. Gossen, M., Freundlieb, S., Bender, G., Muller, G., Hillen, W., and Bujard, H. (1995). Transcriptional activation by tetracyclines in mammalian cells. Science *268*, 1766-1769.

5. Hochedlinger, K., Yamada, Y., Beard, C., and Jaenisch, R. (2005). Ectopic expression of Oct-4 blocks progenitor-cell differentiation and causes dysplasia in epithelial tissues. Cell *121*, 465-477.

6. Hurt, J.A., Thibodeau, S.A., Hirsh, A.S., Pabo, C.O., and Joung, J.K. (2003). Highly specific zinc finger proteins obtained by directed domain shuffling and cell-based selection. Proc Natl Acad Sci U S A *100*, 12271-12276.

7. Korndorfer, I.P., and Skerra, A. (2002). Improved affinity of engineered streptavidin for the Strep-tag II peptide is due to a fixed open conformation of the lid-like loop at the binding site. Protein Sci *11*, 883-893.

8. Seibler, J., Kleinridders, A., Kuter-Luks, B., Niehaves, S., Bruning, J.C., and Schwenk, F. (2007). Reversible gene knockdown in mice using a tight, inducible shRNA expression system. Nucleic Acids Res *35*, e54.

9. Wilson, I.A., Niman, H.L., Houghten, R.A., Cherenson, A.R., Connolly, M.L., and Lerner, R.A. (1984). The structure of an antigenic determinant in a protein. Cell *37*, 767-778.

10. Yan, X., and Sun, X. (2009). Expression of putative zinc-finger protein lcn61 gene in lymphocystis disease virus China (LCDV-cn) genome. Chinese Journal of Oceanology and Limnology *27*, 337-341.

CHAPTER 5:

CONCLUSIONS AND SUMMARY

This thesis has focused on the unique eutherian gene, *ZSCAN5B*, its primate-specific paralogs, *ZSCAN5A, ZSCAN5C and ZSCAN5D.* In Chapter 2, I have shown that the genes duplicated in early primate history and diverged quickly in their DNA-binding "fingerprint" patterns, but have been conserved in sequence since that early divergence. While I could find no tissue or cell line in which ZSCAN5C is expressed, the high level of conservation of this paralog in primates suggests that it likely does have a function in limited cell types, tissues, or conditions that were not explored here. The three primate paralogs, *ZSCAN5A*, B, and D, that are expressed show different human tissue-specific patterns of expression, but overlap in many tissues and are robustly expressed in tissue with rapidly dividing cell populations. Using *in situ* hybridization, I showed that *ZSCAN5B* is expressed within the dividing cell population in those tissues, and that the mouse homolog, *Zscan5b,* is widely expressed in dividing cells in embryogenesis.

Using chromatin immunoprecipitation (ChIP-seq), I also showed that all three expressed ZSCAN5 proteins bind preferentially to Pol III transcription units, especially tDNAs, and related repetitive elements. ZSCAN5B binds with high specificity to a particular subset of tDNA sequences in two human cell lines, and Zscan5b binds to the same conserved subset in mouse tissues. Primate-specific paralogs ZSCAN5A and ZSCAN5D also occupy tDNAs and other RNA Polymerase III (Pol III) promoters, but ZSCAN5D showed a special preference for LINE2 and MIR SINE-associated extra-TFIIIC (ETC) sites throughout the genome. Combining ChIP-seq and RNA-seq data after *ZSCAN5* gene knockdown, I provided evidence to suggest that ZSCAN5B protein binding not only regulates expression of Pol III genes, but also affects the expression of neighboring Pol II genes implicated in ribosome biogenesis, tRNA processing, cell cycle progression and cellular differentiation.

In Chapter 3, I further investigated the hypothesis that human ZSCAN5s are involved in cell cycle progression by manipulating the level of expression in mammalian cells. Consistent with an essential role in cell cycle progression *ZSCAN5A* knockdown

led to the accumulation of cells in the mitotic M-phase and the appearance of aneuploid cells. In Chapter 4, I describe the development tools to further investigate the biological roles of the ancestral *Zscan5b* in the mouse model system, including inducible overexpression and shRNA plasmid constructs. These constructs have been used to establish stable cell lines and importantly, transgenic mouse lines that will allow future investigations of the developmental roles of *Zscan5b* and the effects of gene knockdown or over-expression on mouse health, fertility and behavior.

Taken together, these studies suggest that ZSCAN5B arose in eutherian species to modulate the very ancient process of tRNA synthesis, and with it very likely, the most ancient mechanism of three-dimensional chromatin organization. The three additional family members added in early primate lineages have retained a preference for Pol III transcript binding, and our data indicate that the dominantly expressed human paralog, ZSCAN5A, acts cooperatively with ZSCAN5B. Together these proteins appear to impact cell division and differentiation in a wide variety of developing and adult cells.

As we highlight in Chapter 2, one of the direct targets of ZSCAN5A and ZSCAN5B cooperative binding is *RMRP,* a gene encoding the RNA component of RNase P, and many ZSCAN5-regulated genes– from ribosome biogenesis and tRNA processing to cell cycle regulation – could possibly be explained as downstream effects on this single Pol III target gene. However dysregulation of tRNA expression, or effects on the flanking Pol II genes, almost certainly play a role.

With present data, we cannot be certain about the mechanisms through which ZSCAN5 proteins regulate either the Pol III genes themselves, or the flanking Pol II genes. Published data from other groups suggest that the Pol II genes may be dysregulated simply as a secondary effect of the modulation of Pol III transcripts. And it seems clear that modulating Pol III gene activity should have direct effects on three-dimensional chromatin structure. Future studies, including the application of techniques to measure changes in chromatin architecture, as well as co-immunoprecipitation (co-IP) analysis to query potential binding partners of the proteins throughout the cell cycle, will allow us to gain mechanistic information about the ZSCAN5 genes and their regulatory functions in mammalian cells.

APPENDIX: SUPPLEMENTAL TABLES AND FIGURES

Table A.1. Analyses of affected ZSCAN5 family transcript expression level after specific ZSCAN5 gene ablation

qRT-PCR was done to measure the level of cross-reactivity among ZSCAN5 family after individual siRNA treatments (see methods). Standard deviations were measured from experimental triplicates.

	ZSCAN5A (%)	STDEV		ZSCAN5B (%)	STDEV		ZSCAN5D (%)	STDEV
HEK ZSCAN5A si4	28.4	5.5	HEK ZSCAN5A si4	148.7	13.4	HEK ZSCAN5A si4	59.0	9.1
HEK ZSCAN5A si5	27.4	3.0	HEK ZSCAN5A si5	95.3	29.8	HEK ZSCAN5A si5	46.4	4.6
HEK ZSCAN5B si1	106.2	3.9	HEK ZSCAN5B si1	23.1	6.0	HEK ZSCAN5B si1	73.7	13.4
HEK ZSCAN5D si2	92.6	1.9	HEK ZSCAN5D si2	147.2	9.5	HEK ZSCAN5D si2	32.4	16.2
HEK Sc Ctrl	100.0	8.7	HEK Sc Ctrl	100.0	7.9	HEK Sc Ctrl	100.0	17.5

Table A.2. Combined data from RNA-seq and ChIP-seq experiments merged and organized by chromosome location

ChIP peaks were determined using MACS software, except for 5B_HEK293-Homer peaks which were determined using HOMER software, as described in Methods. Overlap with Oler et al. 2010-Multiple cell type (=OM) or Moqtaderi et al., 2010 RPC155 (R), TFIIIC(T) or ETC(ETC) peaks. Original data table contains more than 4,800 rows. Due to page limitation, only the first page is shown below.

peak type (overlaps)	EXPT	Location: hg19			ChIP repeat overlaps	EF	FDR	DEG	log2 FC	FPKM CTRL	FPKM EXPT	p-value	q-value
.	5Asi4si5	chr1	134772	140566				LOC729737	-1.23578	1.43129	0.607742	5.00E-05	0.00349381
.	5Asi4si5	chr1	700244	714068				LOC100288069	-0.887029	6.46097	3.49362	0.0002	0.00988148
.	5Bsi1	chr1	700244	714068				LOC100288069	-0.598385	6.5542	4.32901	0.0709	na
.	5Bsi1	chr1	934341	935552				HES4	0.690405	203.907	329.053	0.01465	na
.	5Asi4si5	chr1	935502	991499				AGRN	-0.785693	49.7257	28.8445	5.00E-05	0.00349381
ETC	5D_BEWO	chr1	1118942	1119142	.	16.54	0.08						
.	5Asi4si5	chr1	1215815	1227409				SCNN1D	-1.05666	2.2621	1.08749	5.00E-05	0.00349381
.	5Bsi1	chr1	1243993	1260067				PUSL1	0.676556	32.6758	52.2264	0.19795	na
.	5Bsi1	chr1	1270657	1284492				DVL1	0.65568	43.6658	68.7893	0.02565	na
.	5Bsi1	chr1	1288070	1293915				MXRA8	0.700211	2.77305	4.50549	0.04455	na
.	5Bsi1	chr1	1309109	1310818				AURKAIP1	0.910008	176.459	331.577	0.00155	na
.	5Bsi1	chr1	1407163	1431582				ATAD3B	0.815597	9.57606	16.8541	0.0078	na
.	5Bsi1	chr1	1447522	1470067				ATAD3A	0.6632	32.7999	51.9416	0.0192	na
.	5D_BEWO	chr1	1510205	1510405	.	17.89	0.09						
.	5Bsi1	chr1	1533387	1535476				C1orf233	0.864771	12.1122	22.0569	0.0056	na
.	5Asi4si5	chr1	1656276	1677438				SLC35F2	-0.631626	6.42812	4.14902	0.00475	0.0759276
.	5A_BEWO	chr1	1663917	1664117	.	14.85	0.05						
.	5A_BEWO	chr1	1663917	1664117	.	14.85	0.05						
.	5Bsi1	chr1	1981908	2139172				C1orf86	0.944116	49.6105	95.4511	0.032	na
.	5A_BEWO	chr1	2171531	2171731	.	14.85	0.07						
ETC	5D_BEWO	chr1	2478126	2478326	.	30.12	0						
ETC	5D_BEWO	chr1	3370779	3370979	.	15.06	0.15						
.	5Bsi1	chr1	3689351	3692546				LOC388588	1.02345	4.31219	8.76572	0.01305	na
.	5D_BEWO	chr1	3795780	3795980	L1MEe, MLT1A	43.6	0						
.	5Bsi1	chr1	6324331	6453826				ACOT7	0.729162	77.7012	128.803	0.00935	na
.	5Asi4si5	chr1	6581406	6614658				NOL9	-0.686414	6.81162	4.23271	0.00015	0.00803321
.	5Asi4si5	chr1	6673755	6684093				PHF13	-0.603957	19.9063	13.0973	0.0015	0.0380813
.	5D_BEWO	chr1	8184950	8185150	L1MC4, AluSx,	22.99	0.07						
.	5Asi4si5	chr1	8412463	8877699				RERE	-0.858236	12.0956	6.67226	5.00E-05	0.00349381
.	5Bsi1	chr1	8921058	8939943				ENO1-AS1	1.01377	2.31941	4.6833	0.25175	na
.	5D_BEWO	chr1	9488964	9489164	.	19.55	0.09						
.	5Bsi1	chr1	10003485	10045556				NMNAT1	-0.832624	2.82553	1.58655	0.0089	na
.	5D_BEWO	chr1	10143247	10143447	L1PA8, L2a,	15.06	0.1						
.	5D_BEWO	chr1	10969804	10970004	MIRb, GC_rich,	24.06	0.08						
.	5D_BEWO	chr1	11120080	11120280	G-rich	21.05	0						
.	5A_BEWO	chr1	11175217	11175417	.	14.85	0.04						
.	5A_BEWO	chr1	11376060	11376260	.	14.85	0.05						
.	5Bsi1	chr1	11708417	11723384				FBXO2	0.971619	3.68441	7.22528	0.12415	na
.	5Bsi1	chr1	11724149	11734409				FBXO6	0.65001	4.87275	7.64622	0.0466	na
.	5Bsi1	chr1	11796141	11810828				AGTRAP	0.643439	26.9404	42.0822	0.02855	na

Table A.3. Densitometric analyses of affected ZSCAN5 family protein expression level after specific ZSCAN5 gene ablation

After siRNA knockdown, resulting Western Blot was processed and analyzed with ImageJ software to quantitatively evaluate the knock down rates in protein level. Signal intensity values of internal control (TBP) of each sample was used to normalize and calculate the % expression after gene ablation.

	Samples density	TBP control density	Normalization against TBP control	% Expression
ZSCAN5A si	1324.77	21704.903	0.061035518	15.87
Sc ctrl	9050.255	23533.731	0.384565244	
ZSCAN5B si	3401.79	20085.388	0.169366407	32.97
Sc ctrl	9608.154	18705.702	0.513648405	
ZSCAN5D si	1718.841	25027.711	0.068677515	19.93
Sc ctrl	8346.912	24229.468	0.344494233	

Table A.4. Element repeat overlap summary

Repeat Family	Repeat Element	Element (Family)	ZSCAN5B_BeWo	ZSCAN5B_HEK	ZSCAN5A_BeWo_ef1a	ZSCAN5D_BeWo	ETC	RPC155	TFIIIC
scRNA	HY5	HY5 (scRNA)	0.00E+00 ↑					1.29E-110 ↑	
tRNA	tRNA-Ala-GCA	tRNA-Ala-GCA (tRNA)	0.00E+00 ↑		0.00E+00 ↑			0.00E+00 ↑	0.00E+00 ↑
tRNA	tRNA-Ala-GCG	tRNA-Ala-GCG (tRNA)	0.00E+00 ↑	0.00E+00 ↑	0.00E+00 ↑	0.00E+00 ↑		0.00E+00 ↑	0.00E+00 ↑
tRNA	tRNA-Ala-GCY	tRNA-Ala-GCY (tRNA)	0.00E+00 ↑	0.00E+00 ↑	3.86E-38 ↑			0.00E+00 ↑	0.00E+00 ↑
tRNA	tRNA-Ala-GCY_	tRNA-Ala-GCY_ (tRNA)	0.00E+00 ↑	0.00E+00 ↑				0.00E+00 ↑	0.00E+00 ↑
tRNA	tRNA-Arg-AGA	tRNA-Arg-AGA (tRNA)	0.00E+00 ↑	0.00E+00 ↑	0.00E+00 ↑	5.15E-147 ↑		0.00E+00 ↑	1.69E-111 ↑
tRNA	tRNA-Arg-AGG	tRNA-Arg-AGG (tRNA)	0.00E+00 ↑	0.00E+00 ↑	0.00E+00 ↑	0.00E+00 ↑		0.00E+00 ↑	2.13E-293 ↑
tRNA	tRNA-Arg-CGA	tRNA-Arg-CGA (tRNA)	0.00E+00 ↑	0.00E+00 ↑				0.00E+00 ↑	0.00E+00 ↑
tRNA	tRNA-Arg-CGA_	tRNA-Arg-CGA_ (tRNA)	0.00E+00 ↑	0.00E+00 ↑		0.00E+00 ↑		0.00E+00 ↑	1.29E-110 ↑
tRNA	tRNA-Arg-CGG	tRNA-Arg-CGG (tRNA)	0.00E+00 ↑	0.00E+00 ↑	0.00E+00 ↑	1.29E-110 ↑		0.00E+00 ↑	0.00E+00 ↑
tRNA	tRNA-Arg-CGY_	tRNA-Arg-CGY_ (tRNA)	0.00E+00 ↑	0.00E+00 ↑	0.00E+00 ↑	1.29E-110 ↑		0.00E+00 ↑	1.01E-177 ↑
tRNA	tRNA-Asn-AAC	tRNA-Asn-AAC (tRNA)	0.00E+00 ↑	1.29E-110 ↑	0.00E+00 ↑	2.28E-15 ↑		0.00E+00 ↑	0.00E+00 ↑
tRNA	tRNA-Asp-GAY	tRNA-Asp-GAY (tRNA)	0.00E+00 ↑	0.00E+00 ↑	2.75E-56 ↑	4.09E-89 ↑		0.00E+00 ↑	1.02E-231 ↑
tRNA	tRNA-Cys-TGY	tRNA-Cys-TGY (tRNA)	0.00E+00 ↑	0.00E+00 ↑	1.44E-35 ↑	5.01E-08 ↑		0.00E+00 ↑	0.00E+00 ↑
tRNA	tRNA-Gln-CAA	tRNA-Gln-CAA (tRNA)	0.00E+00 ↑	0.00E+00 ↑	0.00E+00 ↑	6.25E-20 ↑		0.00E+00 ↑	9.27E-166 ↑
tRNA	tRNA-Gln-CAG	tRNA-Gln-CAG (tRNA)	0.00E+00 ↑					0.00E+00 ↑	2.58E-223 ↑
tRNA	tRNA-Glu-GAG_	tRNA-Glu-GAG_ (tRNA)	0.00E+00 ↑		0.00E+00 ↑			0.00E+00 ↑	0.00E+00 ↑
tRNA	tRNA-Gly-GGA	tRNA-Gly-GGA (tRNA)	0.00E+00 ↑	0.00E+00 ↑	4.28E-64 ↑	1.25E-74 ↑		2.36E-15 ↑	3.85E-45 ↑
tRNA	tRNA-Gly-GGY	tRNA-Gly-GGY (tRNA)	0.00E+00 ↑	0.00E+00 ↑	0.00E+00 ↑			0.00E+00 ↑	0.00E+00 ↑
tRNA	tRNA-His-CAY_	tRNA-His-CAY_ (tRNA)	0.00E+00 ↑		0.00E+00 ↑	3.86E-38 ↑		0.00E+00 ↑	6.87E-277 ↑
tRNA	tRNA-Ile-ATA	tRNA-Ile-ATA (tRNA)	0.00E+00 ↑	0.00E+00 ↑	3.86E-38 ↑	3.86E-38 ↑		0.00E+00 ↑	6.23E-91 ↑
tRNA	tRNA-Ile-ATT	tRNA-Ile-ATT (tRNA)	0.00E+00 ↑	0.00E+00 ↑	0.00E+00 ↑			0.00E+00 ↑	0.00E+00 ↑
tRNA	tRNA-Leu-CTA	tRNA-Leu-CTA (tRNA)	0.00E+00 ↑	0.00E+00 ↑	0.00E+00 ↑	0.00E+00 ↑		1.29E-110 ↑	0.00E+00 ↑
tRNA	tRNA-Leu-CTA_	tRNA-Leu-CTA_ (tRNA)	0.00E+00 ↑						1.29E-110 ↑
tRNA	tRNA-Leu-CTG	tRNA-Leu-CTG (tRNA)	0.00E+00 ↑	0.00E+00 ↑	1.29E-110 ↑	2.75E-56 ↑		0.00E+00 ↑	1.25E-74 ↑
tRNA	tRNA-Leu-CTY	tRNA-Leu-CTY (tRNA)	0.00E+00 ↑	0.00E+00 ↑	0.00E+00 ↑	3.86E-38 ↑		5.15E-147 ↑	0.00E+00 ↑
tRNA	tRNA-Leu-TTA	tRNA-Leu-TTA (tRNA)	0.00E+00 ↑	0.00E+00 ↑	1.29E-110 ↑			0.00E+00 ↑	1.65E-142 ↑
tRNA	tRNA-Leu-TTG	tRNA-Leu-TTG (tRNA)	0.00E+00 ↑	0.00E+00 ↑	0.00E+00 ↑	1.29E-110 ↑		0.00E+00 ↑	1.25E-74 ↑
tRNA	tRNA-Lys-AAA	tRNA-Lys-AAA (tRNA)	0.00E+00 ↑	0.00E+00 ↑	0.00E+00 ↑			0.00E+00 ↑	1.65E-142 ↑
tRNA	tRNA-Lys-AAG	tRNA-Lys-AAG (tRNA)	0.00E+00 ↑	0.00E+00 ↑	0.00E+00 ↑	2.56E-17 ↑		0.00E+00 ↑	2.58E-223 ↑
tRNA	tRNA-Met	tRNA-Met (tRNA)	0.00E+00 ↑			2.75E-56 ↑		0.00E+00 ↑	9.29E-92 ↑
tRNA	tRNA-Met_	tRNA-Met_ (tRNA)	0.00E+00 ↑	0.00E+00 ↑	0.00E+00 ↑	4.77E-29 ↑		0.00E+00 ↑	1.21E-100 ↑
tRNA	tRNA-Met-i	tRNA-Met-i (tRNA)	0.00E+00 ↑	0.00E+00 ↑	0.00E+00 ↑	1.29E-110 ↑		0.00E+00 ↑	0.00E+00 ↑
tRNA	tRNA-Phe-TTY	tRNA-Phe-TTY (tRNA)	0.00E+00 ↑	0.00E+00 ↑	0.00E+00 ↑	1.29E-23 ↑		0.00E+00 ↑	0.00E+00 ↑
tRNA	tRNA-Pro-CCA	tRNA-Pro-CCA (tRNA)	0.00E+00 ↑	0.00E+00 ↑		0.00E+00 ↑		0.00E+00 ↑	0.00E+00 ↑
tRNA	tRNA-Pro-CCG	tRNA-Pro-CCG (tRNA)	0.00E+00 ↑	0.00E+00 ↑	0.00E+00 ↑	0.00E+00 ↑		0.00E+00 ↑	0.00E+00 ↑
tRNA	tRNA-Pro-CCY	tRNA-Pro-CCY (tRNA)	0.00E+00 ↑		0.00E+00 ↑	0.00E+00 ↑		0.00E+00 ↑	2.13E-293 ↑
tRNA	tRNA-Ser-AGY	tRNA-Ser-AGY (tRNA)	0.00E+00 ↑	0.00E+00 ↑	0.00E+00 ↑	0.00E+00 ↑		0.00E+00 ↑	0.00E+00 ↑
tRNA	tRNA-Ser-TCA	tRNA-Ser-TCA (tRNA)	0.00E+00 ↑	0.00E+00 ↑	0.00E+00 ↑	0.00E+00 ↑		0.00E+00 ↑	0.00E+00 ↑
tRNA	tRNA-Ser-TCA_	tRNA-Ser-TCA_ (tRNA)	0.00E+00 ↑	0.00E+00 ↑	0.00E+00 ↑				0.00E+00 ↑
tRNA	tRNA-Ser-TCG	tRNA-Ser-TCG (tRNA)	0.00E+00 ↑	0.00E+00 ↑	0.00E+00 ↑			0.00E+00 ↑	0.00E+00 ↑
tRNA	tRNA-Ser-TCY	tRNA-Ser-TCY (tRNA)	0.00E+00 ↑	0.00E+00 ↑	0.00E+00 ↑			0.00E+00 ↑	0.00E+00 ↑
tRNA	tRNA-Thr-ACA	tRNA-Thr-ACA (tRNA)	0.00E+00 ↑	0.00E+00 ↑	0.00E+00 ↑			0.00E+00 ↑	1.65E-142 ↑
tRNA	tRNA-Thr-ACG	tRNA-Thr-ACG (tRNA)	0.00E+00 ↑		0.00E+00 ↑	1.29E-110 ↑		0.00E+00 ↑	0.00E+00 ↑
tRNA	tRNA-Thr-ACG_	tRNA-Thr-ACG_ (tRNA)	0.00E+00 ↑	0.00E+00 ↑				0.00E+00 ↑	0.00E+00 ↑
tRNA	tRNA-Thr-ACY	tRNA-Thr-ACY (tRNA)	0.00E+00 ↑	0.00E+00 ↑	1.29E-110 ↑	0.00E+00 ↑		0.00E+00 ↑	1.40E-23 ↑
tRNA	tRNA-Thr-ACY_	tRNA-Thr-ACY_ (tRNA)	0.00E+00 ↑	0.00E+00 ↑	1.29E-110 ↑			0.00E+00 ↑	0.00E+00 ↑
tRNA	tRNA-Trp-TGG	tRNA-Trp-TGG (tRNA)	0.00E+00 ↑	0.00E+00 ↑	2.44E-219 ↑	0.00E+00 ↑		0.00E+00 ↑	2.13E-293 ↑
tRNA	tRNA-Tyr-TAC	tRNA-Tyr-TAC (tRNA)	0.00E+00 ↑	0.00E+00 ↑	0.00E+00 ↑	0.00E+00 ↑		0.00E+00 ↑	0.00E+00 ↑
tRNA	tRNA-Val-GTA	tRNA-Val-GTA (tRNA)	0.00E+00 ↑	0.00E+00 ↑	0.00E+00 ↑			1.29E-110 ↑	0.00E+00 ↑
tRNA	tRNA-Val-GTG	tRNA-Val-GTG (tRNA)	0.00E+00 ↑	0.00E+00 ↑	0.00E+00 ↑	4.09E-89 ↑		0.00E+00 ↑	0.00E+00 ↑
tRNA	tRNA-Val-GTY	tRNA-Val-GTY (tRNA)	0.00E+00 ↑	0.00E+00 ↑		0.00E+00 ↑		0.00E+00 ↑	0.00E+00 ↑
tRNA	tRNA-Glu-GAA	tRNA-Glu-GAA (tRNA)	1.29E-110 ↑		4.77E-29 ↑			0.00E+00 ↑	1.68E-95 ↑
RNA	7SK	7SK (RNA)	6.25E-20 ↑					1.46E-02 ↑	
ERV1	MER66C	MER66C (ERV1)	1.35E-12 ↑						
Simple_repeat	(CGGGG)n	(CGGGG)n (Simple_repeat)	1.35E-12 ↑			3.13E-154 ↑	9.67E-44 ↑		1.15E-82 ↑
Simple_repeat	(CCCCG)n	(CCCCG)n (Simple_repeat)	1.36E-11 ↑			1.91E-08 ↑	3.20E-32 ↑		1.23E-151 ↑
hAT-Charlie	MER113A	MER113A (hAT-Charlie)	4.91E-07 ↑						
snRNA	U6	U6 (snRNA)	9.11E-07 ↑		7.37E-03 ↑			2.07E-25 ↑	
ERVL	LTR67B	LTR67B (ERVL)	2.24E-03 ↑						
hAT-Charlie	MER20B	MER20B (hAT-Charlie)	5.18E-03 ↑						
L2	L2a	L2a (L2)	1.25E-02 ↓						3.72E-02 ↓
L1	L1MD3	L1MD3 (L1)	1.47E-02 ↑						4.60E-02 ↓
Alu	AluY	AluY (Alu)	1.55E-02 ↓		1.43E-04 ↓	7.18E-09 ↓	1.41E-02 ↓	2.85E-96 ↑	
MIR	MIRb	MIRb (MIR)	2.45E-02 ↓			1.31E-56 ↑		1.20E-02 ↑	

Table A.4. (cont.)

ERVL	LTR16C	LTR16C (ERVL)	3.82E-02 ↑	1.31E-02 ↑					3.28E-02 ↓
Low_complexity	C-rich	C-rich (Low_complexity)	3.92E-02 ↑		6.30E-25 ↑	7.47E-22 ↑			9.84E-34 ↑
Alu	AluSx	AluSx (Alu)	4.67E-02 ↓	6.30E-03 ↓	9.25E-09 ↓	1.48E-66 ↑	1.10E-52 ↑		2.38E-149 ↑
MIR	MIR3	MIR3 (MIR)	4.97E-02 ↓		3.89E-19 ↑				1.32E-03 ↑
Alu	AluJb	AluJb (Alu)		1.72E-03 ↓	1.03E-09 ↓	6.42E-14 ↑			2.33E-15 ↑
Alu	AluJo	AluJo (Alu)		5.95E-03 ↓	3.31E-05 ↓	2.07E-08 ↑	2.55E-02 ↑		7.71E-10 ↑
Alu	AluJr	AluJr (Alu)		3.29E-04 ↓	3.76E-04 ↓	3.55E-34 ↑	1.73E-03 ↑		2.49E-46 ↑
Alu	AluJr4	AluJr4 (Alu)				5.05E-11 ↑			4.20E-06 ↑
Alu	AluSc	AluSc (Alu)		3.14E-02 ↓			5.40E-06 ↑		3.81E-03 ↑
Alu	AluSc5	AluSc5 (Alu)					2.60E-03 ↑		
Alu	AluSc8	AluSc8 (Alu)					6.62E-07 ↑		3.31E-04 ↑
Alu	AluSg	AluSg (Alu)		2.93E-02 ↓	4.73E-04 ↓	4.56E-46 ↑	7.46E-80 ↑		2.98E-117 ↑
Alu	AluSg4	AluSg4 (Alu)				4.99E-19 ↑	8.57E-40 ↑		4.11E-54 ↑
Alu	AluSg7	AluSg7 (Alu)					2.98E-25 ↑		8.58E-20 ↑
Alu	AluSp	AluSp (Alu)		4.57E-02 ↓	1.77E-04 ↓	6.57E-29 ↑	3.28E-54 ↑		1.96E-72 ↑
Alu	AluSq	AluSq (Alu)		3.28E-02 ↓	1.94E-02 ↓	1.28E-07 ↑	1.50E-06 ↑		2.07E-11 ↑
Alu	AluSq10	AluSq10 (Alu)				4.09E-09 ↑	5.55E-03 ↑		7.01E-14 ↑
Alu	AluSq2	AluSq2 (Alu)		1.14E-02 ↓		7.86E-32 ↑	1.29E-53 ↑		7.10E-81 ↑
Alu	AluSq4	AluSq4 (Alu)				2.58E-06 ↑	5.43E-07 ↑		2.80E-11 ↑
Alu	AluSx1	AluSx1 (Alu)		8.52E-04 ↓	2.03E-08 ↓	4.12E-75 ↑	3.17E-25 ↑		5.64E-158 ↑
Alu	AluSx3	AluSx3 (Alu)		4.98E-02 ↓		2.00E-24 ↑	1.04E-28 ↑		9.23E-56 ↑
Alu	AluSx4	AluSx4 (Alu)				4.72E-09 ↑	1.56E-07 ↑		1.78E-11 ↑
Alu	AluSz	AluSz (Alu)		1.47E-03 ↓	7.07E-08 ↓	3.61E-41 ↑	2.31E-28 ↑		2.40E-103 ↑
Alu	AluSz6	AluSz6 (Alu)		3.88E-02 ↓	9.19E-04 ↓	6.40E-16 ↑	5.80E-03 ↑		9.32E-15 ↑
Alu	AluYa5	AluYa5 (Alu)					3.59E-18 ↑		
Alu	AluYa8	AluYa8 (Alu)					2.68E-06 ↑		
Alu	AluYb8	AluYb8 (Alu)					1.05E-03 ↑		
Alu	AluYc	AluYc (Alu)				4.66E-03 ↑	1.75E-07 ↑		2.29E-03 ↑
Alu	AluYc3	AluYc3 (Alu)					1.51E-02 ↑		
Alu	AluYf4	AluYf4 (Alu)					9.35E-07 ↑		
Alu	AluYk11	AluYk11 (Alu)					1.09E-02 ↑		
Alu	AluYk4	AluYk4 (Alu)					1.21E-04 ↑		
Alu	FLAM_A	FLAM_A (Alu)				4.59E-15 ↑			1.43E-10 ↑
Alu	FLAM_C	FLAM_C (Alu)			4.95E-02 ↓	3.69E-30 ↑	7.34E-06 ↑		6.63E-59 ↑
Alu	FRAM	FRAM (Alu)							1.48E-02 ↑
centr	GSATII	GSATII (centr)				9.96E-19 ↑			1.95E-17 ↑
CR1	L3	L3 (CR1)				5.03E-03 ↓			9.35E-05 ↓
Deu	AmnSINE1	AmnSINE1 (Deu)					2.99E-02 ↑		
e)-TGA (tRNA	tRNA-Se	RNA-SeC(e)-TGA (tRNA)					0.00E+00 ↑		1.29E-110 ↑
ERV1	HERV35I-int	HERV35I-int (ERV1)					3.51E-20 ↑		
ERV1	HERVE_a-int	HERVE_a-int (ERV1)					2.83E-02 ↑		
ERV1	HERVE-int	HERVE-int (ERV1)		1.63E-02 ↑					
ERV1	HUERS-P1-int	HUERS-P1-int (ERV1)					1.30E-08 ↑		
ERV1	LOR1b	LOR1b (ERV1)		3.93E-02 ↑					
ERV1	LTR1	LTR1 (ERV1)		4.07E-03 ↑					
ERV1	LTR10A	LTR10A (ERV1)				7.00E-03 ↑			
ERV1	LTR10B1	LTR10B1 (ERV1)					2.05E-05 ↑		1.16E-06 ↑
ERV1	LTR10C	LTR10C (ERV1)					3.08E-02 ↑		
ERV1	LTR15	LTR15 (ERV1)					2.56E-11 ↑		1.03E-03 ↑
ERV1	LTR19B	LTR19B (ERV1)				5.63E-04 ↑	2.93E-04 ↑		2.29E-06 ↑
ERV1	LTR1C	LTR1C (ERV1)					2.47E-04 ↑		
ERV1	LTR23	LTR23 (ERV1)				2.40E-02 ↑			2.73E-03 ↑
ERV1	LTR23-int	LTR23-int (ERV1)					3.84E-05 ↑		
ERV1	LTR24B	LTR24B (ERV1)				1.23E-02 ↑			
ERV1	LTR26E	LTR26E (ERV1)				3.52E-04 ↑			3.43E-04 ↑
ERV1	LTR27	LTR27 (ERV1)		6.49E-04 ↑	2.18E-02 ↑				
ERV1	LTR27B	LTR27B (ERV1)			3.19E-02 ↑				
ERV1	LTR2C	LTR2C (ERV1)		4.84E-10 ↑			1.60E-12 ↑		
ERV1	LTR34	LTR34 (ERV1)					3.39E-03 ↑		
ERV1	LTR37A	LTR37A (ERV1)					3.53E-03 ↑		
ERV1	LTR38C	LTR38C (ERV1)							4.88E-03 ↑
ERV1	LTR45C	LTR45C (ERV1)				5.52E-03 ↑			
ERV1	LTR46	LTR46 (ERV1)			1.08E-04 ↑				
ERV1	LTR51	LTR51 (ERV1)					8.68E-08 ↑		
ERV1	LTR54	LTR54 (ERV1)					4.28E-02 ↑		
ERV1	LTR61	LTR61 (ERV1)				3.80E-05 ↑	5.10E-05 ↑		3.88E-02 ↑

Table A.4. (cont.)

ERV1	LTR6A	LTR6A (ERV1)					3.90E-03 ↑				
ERV1	LTR6B	LTR6B (ERV1)						1.44E-05 ↑	3.67E-02 ↑		
ERV1	LTR71A	LTR71A (ERV1)							1.46E-02 ↑		
ERV1	LTR71B	LTR71B (ERV1)							4.44E-02 ↑		
ERV1	LTR73	LTR73 (ERV1)						1.08E-04 ↑			
ERV1	MER110	MER110 (ERV1)					2.95E-02 ↑		2.96E-02 ↑		
ERV1	MER31A	MER31A (ERV1)			1.36E-04 ↑						
ERV1	MER34D	MER34D (ERV1)						2.77E-11 ↑			
ERV1	MER41-int	MER41-int (ERV1)						8.73E-49 ↑			
ERV1	MER41A	MER41A (ERV1)						2.72E-03 ↑			
ERV1	MER41B	MER41B (ERV1)					3.14E-05 ↑	3.10E-04 ↑	1.94E-07 ↑		
ERV1	MER41E	MER41E (ERV1)							2.47E-03 ↑		
ERV1	MER41G	MER41G (ERV1)			1.35E-12 ↑						
ERV1	MER48	MER48 (ERV1)					1.08E-04 ↑		2.06E-02 ↑		
ERV1	MER4A1	MER4A1 (ERV1)						7.54E-12 ↑			
ERV1	MER4D0	MER4D0 (ERV1)						3.15E-03 ↑			
ERV1	MER50	MER50 (ERV1)			2.70E-02 ↑			4.87E-03 ↑			
ERV1	MER51-int	MER51-int (ERV1)						4.85E-02 ↑			
ERV1	MER51C	MER51C (ERV1)			4.25E-06 ↑						
ERV1	MER51D	MER51D (ERV1)						4.31E-06 ↑	5.16E-03 ↑		
ERV1	MER52-int	MER52-int (ERV1)			2.84E-02 ↑						
ERV1	MER57-int	MER57-int (ERV1)						4.53E-19 ↑			
ERV1	MER57E3	MER57E3 (ERV1)					9.97E-06 ↑		2.48E-02 ↑		
ERV1	MER61A	MER61A (ERV1)							2.85E-03 ↑		
ERV1	MER61B	MER61B (ERV1)					1.35E-02 ↑				
ERV1	MER61E	MER61E (ERV1)					3.95E-03 ↑		1.17E-07 ↑		
ERV1	MER65D	MER65D (ERV1)						4.41E-11 ↑			
ERV1	MER67A	MER67A (ERV1)						9.59E-04 ↑	1.95E-03 ↑		
ERV1	MER83B-int	MER83B-int (ERV1)			2.04E-04 ↑			1.60E-10 ↑			
ERV1	MER84	MER84 (ERV1)					5.18E-03 ↑				
ERV1	MER92B	MER92B (ERV1)						3.95E-02 ↑			
ERV1	MER92C	MER92C (ERV1)						4.94E-07 ↑	2.72E-02 ↑		
ERV1	PABL_A-int	PABL_A-int (ERV1)						2.32E-02 ↑			
ERV1	PRIMAX-int	PRIMAX-int (ERV1)							1.23E-02 ↑		
ERVK	HERVK14-int	HERVK14-int (ERVK)						6.64E-03 ↑			
ERVK	HERVK22-int	HERVK22-int (ERVK)			2.39E-02 ↑						
ERVK	HERVKC4-int	HERVKC4-int (ERVK)							2.88E-03 ↑		
ERVK	LTR13	LTR13 (ERVK)			1.80E-02 ↑						
ERVK	LTR13_	LTR13_ (ERVK)						4.77E-29 ↑			
ERVK	LTR14A	LTR14A (ERVK)						2.36E-15 ↑			
ERVL	ERV3-16A3_I-int	ERV3-16A3_I-int (ERVL)							5.16E-03 ↓		
ERVL	ERV3-16A3_LTR	ERV3-16A3_LTR (ERVL)			5.86E-07 ↑						
ERVL	HERVL40-int	HERVL40-int (ERVL)			2.89E-02 ↑						
ERVL	LTR16B	LTR16B (ERVL)			3.95E-03 ↑						
ERVL	LTR16B1	LTR16B1 (ERVL)			4.02E-02 ↑						
ERVL	LTR18B	LTR18B (ERVL)						6.21E-03 ↑			
ERVL	LTR33	LTR33 (ERVL)			9.81E-03 ↑						
ERVL	LTR40b	LTR40b (ERVL)			2.52E-02 ↑						
ERVL	LTR40c	LTR40c (ERVL)						8.17E-03 ↑			
ERVL	LTR41B	LTR41B (ERVL)	7.99E-14 ↑		1.92E-02 ↑						
ERVL	LTR42	LTR42 (ERVL)			2.05E-05 ↑						
ERVL	LTR47A	LTR47A (ERVL)						2.67E-02 ↑			
ERVL	LTR50	LTR50 (ERVL)						2.12E-02 ↑	4.66E-04 ↑		
ERVL	LTR52	LTR52 (ERVL)						2.76E-04 ↑	1.08E-02 ↑		
ERVL	LTR62	LTR62 (ERVL)						1.51E-03 ↑			
ERVL	LTR84b	LTR84b (ERVL)						5.30E-04 ↑			
ERVL	MER54B	MER54B (ERVL)			1.08E-03 ↑		7.00E-03 ↑				
ERVL	MER68B	MER68B (ERVL)			1.83E-03 ↑						
ERVL	MER70C	MER70C (ERVL)			2.05E-05 ↑						
ERVL	MER74C	MER74C (ERVL)						1.61E-06 ↑			
ERVL	MER88	MER88 (ERVL)			1.36E-11 ↑						
ERVL-MaLR	MLT1A	MLT1A (ERVL-MaLR)							1.01E-02 ↓		
ERVL-MaLR	MLT1A-int	MLT1A-int (ERVL-MaLR)			1.33E-03 ↑						
ERVL-MaLR	MLT1A0	MLT1A0 (ERVL-MaLR)							2.49E-04 ↓		
ERVL-MaLR	MLT1B	MLT1B (ERVL-MaLR)							2.93E-03 ↓		
ERVL-MaLR	MLT1C	MLT1C (ERVL-MaLR)					3.32E-02 ↓		6.40E-04 ↓		

Table A.4. (cont.)

ERVL-MaLR	MLT1D	MLT1D (ERVL-MaLR)						1.54E-02 ↓		6.60E-04 ↓
ERVL-MaLR	MLT1E	MLT1E (ERVL-MaLR)			1.39E-02 ↑				4.48E-02 ↑	
ERVL-MaLR	MLT1F2	MLT1F2 (ERVL-MaLR)								2.83E-02 ↓
ERVL-MaLR	MLT1G3	MLT1G3 (ERVL-MaLR)			5.26E-03 ↑					
ERVL-MaLR	MLT1I	MLT1I (ERVL-MaLR)								1.89E-02 ↓
ERVL-MaLR	MLT1J	MLT1J (ERVL-MaLR)	3.47E-02 ↑							8.22E-03 ↓
ERVL-MaLR	MLT1K	MLT1K (ERVL-MaLR)				4.97E-02 ↓				2.80E-03 ↓
ERVL-MaLR	MLT1L	MLT1L (ERVL-MaLR)								1.05E-02 ↓
ERVL-MaLR	MSTA	MSTA (ERVL-MaLR)			4.02E-02 ↓			1.87E-02 ↓	4.41E-02 ↓	1.28E-04 ↓
ERVL-MaLR	MSTB	MSTB (ERVL-MaLR)								9.96E-03 ↓
ERVL-MaLR	THE1B	THE1B (ERVL-MaLR)			2.69E-02 ↓			1.04E-02 ↓		1.01E-05 ↓
ERVL-MaLR	THE1B-int	THE1B-int (ERVL-MaLR)								3.08E-03 ↓
ERVL-MaLR	THE1C	THE1C (ERVL-MaLR)	1.41E-02 ↑							4.65E-03 ↓
ERVL-MaLR	THE1D	THE1D (ERVL-MaLR)								1.38E-03 ↓
Gypsy	LTR81A	LTR81A (Gypsy)			5.52E-03 ↑					
Gypsy	LTR81C	LTR81C (Gypsy)			2.89E-03 ↑					
Gypsy	MamGypLTR1a	MamGypLTR1a (Gypsy)			1.51E-02 ↑					
Gypsy	MamGypLTR3	MamGypLTR3 (Gypsy)			4.23E-03 ↑					
Gypsy?	LTR85b	LTR85b (Gypsy?)			2.74E-02 ↑					
Gypsy?	LTR88b	LTR88b (Gypsy?)			3.93E-05 ↑					
hAT	MamRep1894	MamRep1894 (hAT)				3.68E-03 ↑				
hAT	MER96	MER96 (hAT)			8.20E-03 ↑				1.61E-05 ↑	
hAT-Blackjack	MER81	MER81 (hAT-Blackjack)			3.87E-03 ↑					
hAT-Charlie	Charlie10b	Charlie10b (hAT-Charlie)							9.16E-07 ↑	5.69E-06 ↑
hAT-Charlie	Charlie12	Charlie12 (hAT-Charlie)								2.36E-15 ↑
hAT-Charlie	Charlie17a	Charlie17a (hAT-Charlie)			7.00E-03 ↑					
hAT-Charlie	Charlie1a	Charlie1a (hAT-Charlie)							2.06E-02 ↑	
hAT-Charlie	MER102c	MER102c (hAT-Charlie)				6.97E-03 ↑				
hAT-Charlie	MER106B	MER106B (hAT-Charlie)				3.33E-02 ↑				
hAT-Charlie	MER113	MER113 (hAT-Charlie)	2.83E-05 ↑						2.49E-03 ↑	
hAT-Charlie	MER20	MER20 (hAT-Charlie)								4.47E-02 ↓
hAT-Charlie	MER3	MER3 (hAT-Charlie)	7.81E-03 ↑							1.73E-02 ↓
hAT-Charlie	MER58A	MER58A (hAT-Charlie)								3.94E-03 ↓
hAT-Charlie	MER58B	MER58B (hAT-Charlie)								3.76E-02 ↓
hAT-Charlie	MER5A	MER5A (hAT-Charlie)						3.99E-02 ↓		1.73E-02 ↓
hAT-Charlie	MER5B	MER5B (hAT-Charlie)				3.93E-02 ↓		4.24E-02 ↓		
hAT-Tip100	Arthur1A	Arthur1A (hAT-Tip100)							1.64E-02 ↑	
hAT-Tip100	FordPrefect_a	ordPrefect_a (hAT-Tip100)			1.17E-07 ↑					
hAT-Tip100	MER45C	MER45C (hAT-Tip100)			1.31E-05 ↑					
hAT-Tip100	MER91A	MER91A (hAT-Tip100)				4.30E-02 ↑				
hAT-Tip100	MER91B	MER91B (hAT-Tip100)								4.50E-03 ↑
hAT-Tip100	MER91C	MER91C (hAT-Tip100)			1.24E-02 ↑					
hAT-Tip100	ORSL	ORSL (hAT-Tip100)			1.13E-02 ↑					
L1	HAL1	HAL1 (L1)						5.30E-03 ↓		1.06E-05 ↓
L1	L1M1	L1M1 (L1)							3.41E-02 ↓	2.02E-04 ↓
L1	L1M2a	L1M2a (L1)			2.20E-02 ↑					
L1	L1M2b	L1M2b (L1)				2.23E-03 ↑				
L1	L1M2c	L1M2c (L1)			4.24E-03 ↑					
L1	L1M3	L1M3 (L1)								1.71E-02 ↓
L1	L1M3b	L1M3b (L1)			1.63E-02 ↑					
L1	L1M4	L1M4 (L1)			3.22E-02 ↓					1.31E-03 ↓
L1	L1M4c	L1M4c (L1)								2.15E-02 ↓
L1	L1M5	L1M5 (L1)			1.87E-02 ↓			2.24E-04 ↓		2.62E-10 ↓
L1	L1M6	L1M6 (L1)								2.10E-02 ↓
L1	L1MA1	L1MA1 (L1)								4.40E-03 ↓
L1	L1MA2	L1MA2 (L1)			4.27E-02 ↓					
L1	L1MA3	L1MA3 (L1)			2.85E-02 ↓					
L1	L1MA4A	L1MA4A (L1)								1.22E-03 ↓
L1	L1MA8	L1MA8 (L1)								8.71E-04 ↓
L1	L1MA9	L1MA9 (L1)			3.23E-02 ↓					1.60E-05 ↓
L1	L1MB1	L1MB1 (L1)								6.03E-03 ↓
L1	L1MB2	L1MB2 (L1)								1.06E-02 ↓
L1	L1MB3	L1MB3 (L1)			4.02E-02 ↓			1.68E-02 ↓		6.93E-03 ↓
L1	L1MB5	L1MB5 (L1)								2.62E-02 ↓
L1	L1MB7	L1MB7 (L1)			1.40E-02 ↓			3.12E-02 ↓		1.15E-04 ↓
L1	L1MB8	L1MB8 (L1)			3.23E-02 ↓			2.51E-02 ↓		6.56E-03 ↓

Table A.4. (cont.)

L1	L1MC	L1MC (L1)						1.86E-02 ↓
L1	L1MC1	L1MC1 (L1)		4.03E-02 ↓			3.94E-02 ↓	9.45E-05 ↓
L1	L1MC3	L1MC3 (L1)						2.69E-03 ↓
L1	L1MC4	L1MC4 (L1)		2.27E-02 ↓	8.85E-03 ↓			3.59E-05 ↓
L1	L1MC4a	L1MC4a (L1)		1.61E-02 ↓	4.77E-03 ↓			1.12E-03 ↓
L1	L1MC5	L1MC5 (L1)		4.28E-02 ↓				1.36E-02 ↓
L1	L1MD	L1MD (L1)						1.01E-02 ↓
L1	L1MD1	L1MD1 (L1)						7.42E-03 ↓
L1	L1MD2	L1MD2 (L1)						1.01E-03 ↓
L1	L1MDa	L1MDa (L1)						1.12E-02 ↓
L1	L1ME1	L1ME1 (L1)		1.88E-02 ↓				3.28E-08 ↓
L1	L1ME2	L1ME2 (L1)		4.81E-02 ↓				2.11E-03 ↓
L1	L1ME3	L1ME3 (L1)						7.28E-03 ↓
L1	L1ME3A	L1ME3A (L1)						3.22E-03 ↓
L1	L1ME3B	L1ME3B (L1)						1.74E-02 ↓
L1	L1ME3C	L1ME3C (L1)						6.76E-04 ↓
L1	L1ME3E	L1ME3E (L1)						4.85E-02 ↓
L1	L1ME4a	L1ME4a (L1)		1.16E-02 ↓	2.59E-03 ↓	2.67E-03 ↓		3.50E-08 ↓
L1	L1MEa	L1MEa (L1)					8.66E-03 ↑	
L1	L1MEc	L1MEc (L1)		2.57E-02 ↓			2.85E-02 ↓	3.05E-06 ↓
L1	L1MEe	L1MEe (L1)						1.08E-03 ↓
L1	L1MEf	L1MEf (L1)				3.34E-02 ↓		4.78E-04 ↓
L1	L1MEg	L1MEg (L1)		3.43E-02 ↓				5.29E-05 ↓
L1	L1P4e	L1P4e (L1)		1.98E-19 ↑				
L1	L1PA10	L1PA10 (L1)				3.97E-02 ↓		2.43E-04 ↓
L1	L1PA12	L1PA12 (L1)		2.26E-02 ↑				
L1	L1PA15	L1PA15 (L1)		4.13E-02 ↓				
L1	L1PA16	L1PA16 (L1)					1.57E-02 ↓	9.54E-07 ↓
L1	L1PA17	L1PA17 (L1)						1.17E-02 ↓
L1	L1PA2	L1PA2 (L1)		2.51E-02 ↓			2.66E-02 ↓	2.10E-04 ↓
L1	L1PA3	L1PA3 (L1)		3.80E-03 ↓				
L1	L1PA4	L1PA4 (L1)		5.15E-03 ↓				
L1	L1PA5	L1PA5 (L1)		6.54E-03 ↓				
L1	L1PA6	L1PA6 (L1)		2.74E-02 ↓				
L1	L1PA7	L1PA7 (L1)		2.60E-03 ↓			8.99E-04 ↓	
L1	L1PB1	L1PB1 (L1)		1.40E-02 ↓		3.18E-03 ↓		6.18E-07 ↓
L1	L1PREC2	L1PREC2 (L1)						5.26E-04 ↓
L2	L2	L2 (L2)				2.49E-04 ↓		6.22E-08 ↓
L2	L2b	L2b (L2)		2.48E-02 ↑	5.13E-28 ↑	####### ↑	3.17E-02 ↓	4.66E-128 ↑
L2	L2c	L2c (L2)			2.34E-05 ↑	8.20E-54 ↑	2.96E-02 ↓	2.09E-41 ↑
Low_complexity	A-rich	A-rich (Low_complexity)	6.24E-03 ↑					
Low_complexity	AT_rich	T_rich (Low_complexity)		5.80E-06 ↓		1.88E-07 ↓	1.76E-06 ↓	4.84E-19 ↓
Low_complexity	G-rich	G-rich (Low_complexity)			2.09E-29 ↑	2.96E-12 ↑		8.11E-29 ↑
Low_complexity	GC_rich	C_rich (Low_complexity)	3.72E-09 ↑	1.08E-79 ↑	0.00E+00 ↑	####### ↑		0.00E+00 ↑
Low_complexity	polypurine	lypurine (Low_complexity)				2.44E-02 ↑	1.05E-02 ↑	
Low_complexity	polypyrimidine	pyrimidine (Low_complexity)				3.46E-02 ↑		7.04E-05 ↑
Low_complexity	T-rich	T-rich (Low_complexity)						4.39E-02 ↓
LTR	MamRep1527	MamRep1527 (LTR)					1.02E-04 ↑	
LTR?	MER129	MER129 (LTR?)			2.48E-07 ↑			
MIR	MIR	MIR (MIR)			1.53E-32 ↑	2.32E-02 ↑	1.50E-49 ↑	1.15E-02 ↑
MIR	MIRc	MIRc (MIR)			6.28E-71 ↑	4.58E-02 ↑		
rRNA	5S	5S (rRNA)				1.54E-10 ↑	0.00E+00 ↑	0.00E+00 ↑
rRNA	LSU-rRNA_Hsa	LSU-rRNA_Hsa (rRNA)			1.66E-03 ↑			2.78E-100 ↑
rRNA	SSU-rRNA_Hsa	SSU-rRNA_Hsa (rRNA)						3.96E-66 ↑
RTE	L4	L4 (RTE)						1.96E-03 ↓
RTE-BovB	X3_LINE	X3_LINE (RTE-BovB)					3.52E-04 ↑	
Satellite	MSR1	MSR1 (Satellite)			3.39E-03 ↑			
Satellite	SUBTEL_sa	SUBTEL_sa (Satellite)		1.29E-110 ↑		2.75E-56 ↑		2.36E-15 ↑
scRNA	BC200	BC200 (scRNA)					7.99E-14 ↑	9.59E-04 ↑
scRNA	HY1	HY1 (scRNA)						4.59E-02 ↑
scRNA	HY3	HY3 (scRNA)					2.04E-04 ↑	
scRNA	HY4	HY4 (scRNA)					5.01E-08 ↑	
Simple_repeat	(C)n	(C)n (Simple_repeat)						6.64E-06 ↑
Simple_repeat	(CA)n	(CA)n (Simple_repeat)		1.53E-02 ↓			2.81E-02 ↓	
Simple_repeat	(CAA)n	(CAA)n (Simple_repeat)					4.18E-02 ↑	
Simple_repeat	(CAAA)n	(CAAA)n (Simple_repeat)					3.27E-02 ↑	

Table A.4. (cont.)

Simple_repeat	(CAACG)n	CAACG)n (Simple_repeat)						0.00E+00 ↑
Simple_repeat	(CACCC)n	CACCC)n (Simple_repeat)						4.97E-02 ↑
Simple_repeat	(CAG)n	(CAG)n (Simple_repeat)			4.56E-03 ↑	1.27E-07 ↑		6.36E-13 ↑
Simple_repeat	(CAGA)n	(CAGA)n (Simple_repeat)			4.13E-04 ↑			
Simple_repeat	(CAGAGA)n	CAGAGA)n (Simple_repeat)			2.44E-02 ↑		8.66E-03 ↑	
Simple_repeat	(CAGC)n	(CAGC)n (Simple_repeat)					2.50E-07 ↑	
Simple_repeat	(CAGCC)n	CAGCC)n (Simple_repeat)						1.04E-02 ↑
Simple_repeat	(CAGG)n	(CAGG)n (Simple_repeat)				9.41E-11 ↑		1.86E-03 ↑
Simple_repeat	(CAGGG)n	CAGGG)n (Simple_repeat)						9.94E-06 ↑
Simple_repeat	(CAT)n	(CAT)n (Simple_repeat)		1.75E-02 ↑				
Simple_repeat	(CCAA)n	(CCAA)n (Simple_repeat)			9.16E-07 ↑			
Simple_repeat	(CCCCAG)n	CCCCAG)n (Simple_repeat)		4.94E-07 ↑	7.22E-20 ↑			
Simple_repeat	(CCCCCG)n	CCCCCG)n (Simple_repeat)						4.14E-04 ↑
Simple_repeat	(CCCTG)n	CCCTG)n (Simple_repeat)						2.69E-06 ↑
Simple_repeat	(CCG)n	(CCG)n (Simple_repeat)		2.38E-10 ↑	2.05E-67 ↑	3.79E-06 ↑		1.56E-136 ↑
Simple_repeat	(CCGAG)n	CCGAG)n (Simple_repeat)			0.00E+00 ↑			1.40E-23 ↑
Simple_repeat	(CCGCG)n	CCGCG)n (Simple_repeat)						1.60E-06 ↑
Simple_repeat	(CCGGG)n	CCGGG)n (Simple_repeat)			4.77E-29 ↑	2.75E-56 ↑		4.84E-10 ↑
Simple_repeat	(CG)n	(CG)n (Simple_repeat)			9.41E-11 ↑			1.66E-03 ↑
Simple_repeat	(CGAG)n	(CGAG)n (Simple_repeat)						1.98E-09 ↑
Simple_repeat	(CGAGG)n	CGAGG)n (Simple_repeat)						7.99E-14 ↑
Simple_repeat	(CGCGG)n	CGCGG)n (Simple_repeat)						2.28E-45 ↑
Simple_repeat	(CGCTG)n	CGCTG)n (Simple_repeat)						1.29E-110 ↑
Simple_repeat	(CGG)n	(CGG)n (Simple_repeat)		1.09E-03 ↑	2.22E-123 ↑	2.17E-83 ↑		0.00E+00 ↑
Simple_repeat	(CGGA)n	(CGGA)n (Simple_repeat)						5.05E-08 ↑
Simple_repeat	(CGGG)n	(CGGG)n (Simple_repeat)			4.77E-29 ↑			3.07E-26 ↑
Simple_repeat	(CGGGGG)n	CGGGGG)n (Simple_repeat)						1.57E-22 ↑
Simple_repeat	(CGTG)n	(CGTG)n (Simple_repeat)			2.05E-05 ↑			2.96E-02 ↑
Simple_repeat	(CTA)n	(CTA)n (Simple_repeat)				5.61E-04 ↑		
Simple_repeat	(CTCA)n	(CTCA)n (Simple_repeat)			9.59E-04 ↑			
Simple_repeat	(CTCG)n	(CTCG)n (Simple_repeat)						7.99E-14 ↑
Simple_repeat	(CTG)n	(CTG)n (Simple_repeat)			4.55E-03 ↑			
Simple_repeat	(CTGGGG)n	CTGGGG)n (Simple_repeat)		4.30E-31 ↑	9.84E-11 ↑			6.09E-04 ↑
Simple_repeat	(G)n	(G)n (Simple_repeat)						1.44E-35 ↑
Simple_repeat	(GGA)n	(GGA)n (Simple_repeat)		8.56E-05 ↑	6.56E-29 ↑	1.15E-02 ↑		
Simple_repeat	(GGGAA)n	GGGAA)n (Simple_repeat)			6.67E-06 ↑			6.33E-07 ↑
Simple_repeat	(T)n	(T)n (Simple_repeat)	1.36E-03 ↑					
Simple_repeat	(TA)n	(TA)n (Simple_repeat)		7.74E-03 ↓		2.68E-03 ↓	2.76E-03 ↓	1.10E-07 ↓
Simple_repeat	(TCC)n	(TCC)n (Simple_repeat)			2.01E-02 ↑	2.46E-02 ↑		8.72E-08 ↑
Simple_repeat	(TCCC)n	(TCCC)n (Simple_repeat)			3.67E-05 ↑			
Simple_repeat	(TCCCC)n	TCCCC)n (Simple_repeat)			2.05E-03 ↑			4.78E-10 ↑
Simple_repeat	(TCCG)n	(TCCG)n (Simple_repeat)						1.17E-07 ↑
Simple_repeat	(TCG)n	(TCG)n (Simple_repeat)				######## ↑		4.77E-29 ↑
Simple_repeat	(TCTCTG)n	TCTCTG)n (Simple_repeat)			2.41E-05 ↑			
Simple_repeat	(TCTG)n	(TCTG)n (Simple_repeat)		6.71E-09 ↑				
Simple_repeat	(TG)n	(TG)n (Simple_repeat)			3.90E-02 ↓		4.24E-02 ↓	
Simple_repeat	(TGGG)n	(TGGG)n (Simple_repeat)				1.36E-03 ↑		
Simple_repeat	(TTC)n	(TTC)n (Simple_repeat)			3.94E-03 ↑			
Simple_repeat	(TTGGGG)n	TTGGGG)n (Simple_repeat)			1.35E-12 ↑		2.36E-15 ↑	8.56E-05 ↑
Simple_repeat	(TTTG)n	(TTTG)n (Simple_repeat)					2.71E-02 ↑	
Simple_repeat	(TTTTA)n	TTTTA)n (Simple_repeat)		4.41E-02 ↑				
Simple_repeat	(TTTTG)n	TTTTG)n (Simple_repeat)	2.69E-06 ↑					
Simple_repeat	(TTTTTG)n	TTTTTG)n (Simple_repeat)					2.52E-02 ↑	
snRNA	U13	U13 (snRNA)					1.36E-11 ↑	
snRNA	U4	U4 (snRNA)				5.05E-08 ↑		1.08E-04 ↑
srpRNA	7SLRNA	7SLRNA (srpRNA)		1.36E-02 ↑			9.53E-04 ↑	
TcMar-Tc2	Kanga1a	Kanga1a (TcMar-Tc2)					2.93E-02 ↑	
TcMar-Tc2	Kanga1b	Kanga1b (TcMar-Tc2)				3.92E-03 ↑		
TcMar-Tigger	MER127	MER127 (TcMar-Tigger)			2.05E-05 ↑			
TcMar-Tigger	MER44B	MER44B (TcMar-Tigger)					1.09E-02 ↑	
TcMar-Tigger	MER44D	MER44D (TcMar-Tigger)					1.99E-02 ↑	
TcMar-Tigger	MER8	MER8 (TcMar-Tigger)					2.14E-03 ↑	
TcMar-Tigger	Tigger1	Tigger1 (TcMar-Tigger)		2.98E-02 ↓				1.45E-05 ↓
TcMar-Tigger	Tigger4	Tigger4 (TcMar-Tigger)					1.50E-02 ↑	
TcMar-Tigger	Tigger7	Tigger7 (TcMar-Tigger)					4.48E-02 ↑	
TcMar-Tigger	Tigger9b	Tigger9b (TcMar-Tigger)					2.10E-02 ↑	

Table A.4. (cont.)

telo	REP522	REP522 (telo)			8.54E-05	↑					
telo	TAR1	TAR1 (telo)			9.11E-07	↑					
tRNA	tRNA-Gly-GGG	tRNA-Gly-GGG (tRNA)			0.00E+00	↑			0.00E+00	↑	0.00E+00 ↑

Table A.5. Enriched mZscan5b peaks from ChIP-seq in E17.5 fetal placenta

PeakID	Chr	Start	End	Peak Score	Focus Ratio/Region Size	Annotation	Detailed Annotation	Distance to TSS	Gene Name
chr3-3	chr3	90279685	90279907	17.2	0.884	promoter-TSS (NM_026374)	trna-Met-I; promoter-TSS (NM_026374)	-327	Ilf2
chr13-3	chr13	23622130	23622352	15.7	0.932	promoter-TSS (NM_153173)	trna-Met-I; promoter-TSS (NM_153173)	-672	Hist1h4h
chr11-77	chr11	48647830	48648052	15.3	0.926	intron (NM_053166, intron 1 of 6)	tRNA-Ala-GCG|tRNA|tRNA	8301	Trim7
chr14-1	chr14	51709080	51709302	14.3	0.925	Intergenic	tRNA-Tyr-TAC|tRNA|tRNA	-1561	Rnase4
chr1-9	chr1	1.73E+08	1.73E+08	14	0.886	Intergenic	tRNA-Leu-CTG|tRNA|tRNA	19721	Fcgr4
chr13-4	chr13	23529895	23530117	13.3	0.826	Intergenic	tRNA-Ile-ATT|tRNA|tRNA	-7120	C230035I16Rik
chr14-2	chr14	51683489	51683711	13.3	0.891	Intergenic	tRNA-Leu-CTY|tRNA|tRNA	-6683	Rnase12
chr7-2	chr7	1.28E+08	1.28E+08	13.3	0.855	Intergenic	tRNA-Leu-CTA|tRNA|tRNA	-7323	Eef2k
chr14-3	chr14	51684749	51684971	12.8	0.712	Intergenic	tRNA-Thr-ACA|tRNA|tRNA	6257	Olfr750
chr6-4	chr6	1.29E+08	1.29E+08	12.6	0.933	Intergenic	4.5SRNA|scRNA|scRNA	24449	Klrb1c
chr13-5	chr13	23545230	23545452	11.6	0.82	Intergenic	tRNA-Leu-CTG|tRNA|tRNA	12429	Btn1a1
chr5-7	chr5	31190823	31191045	11.4	0.833	promoter-TSS (NR_122117)	promoter-TSS (NR_122117)	-291	Agbl5
chr11-86	chr11	97849474	97849696	11.1	0.806	Intergenic	tRNA-Cys-TGY|tRNA|tRNA	-1825	Plxdc1
chr19-2	chr19	12084054	12084276	11.1	0.828	Intergenic	Intergenic	27065	Olfr1423
chr11-82	chr11	48667035	48667257	10.9	0.904	Intergenic	tRNA-Thr-ACA|tRNA|tRNA	17702	Irgm1
chr11-84	chr11	97850163	97850385	10.7	0.873	TTS (NM_207231)	TTS (NM_207231)	-2514	Plxdc1
chr13-8	chr13	21802791	21803013	10.7	0.733	Intergenic	tRNA-Met-i|tRNA|tRNA	5110	Hist1h2bl
chr13-6	chr13	23613670	23613892	10.7	0.81	Intergenic	tRNA-Ser-AGY|tRNA|tRNA	-9132	Hist1h4h
chr2-11	chr2	1.22E+08	1.22E+08	10.7	0.926	Intergenic	tRNA-His-CAY_|tRNA|tRNA	-6596	Shf
chr13-7	chr13	21291991	21292213	10.4	0.833	Intergenic	tRNA-Leu-TTG|tRNA|tRNA	20302	Trim27
chr11-83	chr11	68854530	68854752	10.4	0.69	Intergenic	Intergenic	-4504	Aurkb
chr11-85	chr11	68875714	68875936	10.2	0.789	TTS (NM_028005)	TTS (NM_028005)	2548	2310047M10Rik
chr7-3	chr7	1.28E+08	1.28E+08	9.69	0.789	intron (NM_025298, intron 1 of 20)	MIR|SINE|MIR	1033	Polr3e
chr16-6	chr16	13437366	13437588	9.69	0.707	Intergenic	tRNA-Thr-ACG|tRNA|tRNA	-12139	Mir193b
chr11-87	chr11	68850928	68851150	9.69	0.677	Intergenic	tRNA-Ser-TCY|tRNA|tRNA	-8106	Aurkb
chr13-9	chr13	21332741	21332963	9.44	0.846	Intergenic	tRNA-Ala-GCA|tRNA|tRNA	51703	Gpx5
chr11-90	chr11	68932204	68932426	9.44	0.792	Intergenic	tRNA-Gly-GGY|tRNA|tRNA	-1640	Hes7
chr11-91	chr11	1.07E+08	1.07E+08	8.96	0.87	Intergenic	Intergenic	-13342	Kpna2
chr6-5	chr6	1.29E+08	1.29E+08	8.96	0.848	Intergenic	4.5SRNA|scRNA|scRNA	27883	Klrb1c
chr5-11	chr5	1.26E+08	1.26E+08	8.72	0.76	Intergenic	tRNA-Ala-GCG|tRNA|tRNA	-19460	Ubc
chr5-10	chr5	1.26E+08	1.26E+08	8.72	0.844	Intergenic	tRNA-Ala-GCG|tRNA|tRNA	-14254	Ubc
chr3-4	chr3	21975245	21975467	8.72	0.692	promoter-TSS (NM_030732)	promoter-TSS (NM_030732)	-218	Tbl1xr1
chr13-12	chr13	21325830	21326052	8.72	0.736	Intergenic	tRNA-Ala-GCY_|tRNA|tRNA	54141	Trim27
chr11-89	chr11	58115584	58115806	8.72	0.661	Intergenic	tRNA-Glu-GAG_|tRNA|tRNA	-4876	Zfp692
chr8-34	chr8	20020269	20020491	8.47	0.726	promoter-TSS (NR_030708).2	promoter-TSS (NR_030708).2	12	2610005L07Rik
chr6-6	chr6	1.29E+08	1.29E+08	8.47	0.804	Intergenic	4.5SRNA|scRNA|scRNA	-14440	Klrb1c
chr5-8	chr5	1.5E+08	1.5E+08	8.47	0.78	Intergenic	tRNA-Asn-AAC|tRNA|tRNA	-36031	Alox5ap
chr18-10	chr18	36961512	36961734	8.47	0.884	ıon-coding (NR_027885, exon 1 of 1	non-coding (NR_027885, exon 1 of 1)	206	Vaultrc5
chr11-94	chr11	68876462	68876684	8.47	0.857	Intergenic	tRNA-Ser-AGY|tRNA|tRNA	3296	2310047M10Rik
chr2-14	chr2	1.22E+08	1.22E+08	8.23	0.83	Intergenic	tRNA-His-CAY_|tRNA|tRNA	-9108	Shf
chr13-11	chr13	23608976	23609198	8.23	0.822	Intergenic	tRNA-Gly-GGY|tRNA|tRNA	-13826	Hist1h4h
chr11-95	chr11	68938542	68938764	7.99	0.745	Intergenic	Intergenic	-1226	Aloxe3
chr15-3	chr15	57807488	57807710	7.99	0.848	Intergenic	tRNA-Met_|tRNA|tRNA	-9859	Fam83a
chr3-6	chr3	96303373	96303595	7.75	0.85	Intergenic	tRNA-Lys-AAG|tRNA|tRNA	-25624	Hfe2
chr13-18	chr13	21261978	21262200	7.75	0.782	Intergenic	tRNA-Leu-CTA_|tRNA|tRNA	-9711	Trim27
chr10-9	chr10	30530859	30531081	7.75	0.872	Intergenic	tRNA-Glu-GAG_|tRNA|tRNA	-8057	Ncoa7
chr13-19	chr13	22028867	22029089	7.75	0.881	Intergenic	tRNA-Ser-TCY|tRNA|tRNA	-7985	Zfp184
chr11-96	chr11	97631080	97631302	7.51	0.778	Intergenic	tRNA-Cys-TGY|tRNA|tRNA	-3264	Cwc25
chr13-20	chr13	22066838	22067060	7.51	0.805	Intergenic	tRNA-Met-i|tRNA|tRNA	-6101	Pom121l2
chr3-5	chr3	96305268	96305490	7.51	0.711	Intergenic	tRNA-Glu-GAG_|tRNA|tRNA	-23729	Hfe2
chr1-145	chr1	1.73E+08	1.73E+08	7.51	0.841	Intergenic	tRNA-Leu-CTG|tRNA|tRNA	-19813	1700009P17Rik
chr3-7	chr3	19528878	19529100	7.26	0.756	Intergenic	Intergenic	-15471	Trim55
chr6-8	chr6	38483896	38484118	7.26	0.865	promoter-TSS (NM_025363)	promoter-TSS (NM_025363)	-854	1110001J03Rik
chr17-11	chr17	23684253	23684475	7.26	0.865	Intergenic	tRNA-Lys-AAG|tRNA|tRNA	16829	Zfp213
chr14-7	chr14	55136451	55136673	7.26	0.721	promoter-TSS (NM_013768)	promoter-TSS (NM_013768)	-255	Prmt5
chr7-4	chr7	1.06E+08	1.06E+08	7.26	0.667	Intergenic	Intergenic	-19565	Wnt11
chr14-6	chr14	51710208	51710430	7.02	0.762	promoter-TSS (NM_001161731)	promoter-TSS (NM_001161731)	-433	Rnase4
chr7-7	chr7	65654595	65654817	7.02	0.805	Intergenic	tRNA-Glu-GAG_|tRNA|tRNA	-258866	Atp10a
chr10-12	chr10	62892150	62892372	7.02	0.886	promoter-TSS (NM_177612)	promoter-TSS (NM_177612)	-585	Ctnna3
chr8-57	chr8	1.14E+08	1.14E+08	7.02	0.744	Intergenic	tRNA-Gly-GGY|tRNA|tRNA	6258	Exosc6
chr7-9	chr7	86611281	86611503	6.78	0.667	promoter-TSS (NM_017462)	promoter-TSS (NM_017462)	-233	Polg
chr3-9	chr3	1.22E+08	1.22E+08	6.54	0.806	TTS (NR_030439)	TTS (NR_030439)	719	Mir760
chr17-15	chr17	23683356	23683578	6.54	0.865	Intergenic	tRNA-Pro-CCA|tRNA|tRNA	17726	Zfp213
chr1-161	chr1	34491554	34491776	6.54	0.833	Intergenic	tRNA-Glu-GAG_|tRNA|tRNA	-1621	Mir5103
chr10-14	chr10	90644764	90644986	6.54	0.707	Intergenic	tRNA-Asp-GAY|tRNA|tRNA	-10511	Tmpo
chr1-127	chr1	1.68E+08	1.68E+08	6.54	0.903	Intergenic	Intergenic	-7805	Mpzl1
chr6-10	chr6	1.29E+08	1.29E+08	6.3	0.721	Intergenic	Intergenic	28437	Klrb1c
chr3-8	chr3	51163231	51163453	6.3	0.778	Intergenic	tRNA-Pro-CCY|tRNA|tRNA	-18776	Elf2
chr1-144	chr1	1.68E+08	1.68E+08	6.3	0.778	Intergenic	tRNA-Pro-CCY|tRNA|tRNA	-7098	Mpzl1
chr1-159	chr1	1.73E+08	1.73E+08	6.3	0.577	Intergenic	Intergenic	-18718	1700009P17Rik
chr11-99	chr11	97563706	97563928	6.05	0.743	promoter-TSS (NM_011971)	promoter-TSS (NM_011971)	-931	Psmb3
chr13-26	chr13	21252530	21252752	6.05	0.765	Intergenic	Intergenic	-17717	Olfr1368
chr10-18	chr10	79711802	79712024	6.05	0.762	promoter-TSS (NM_029272)	promoter-TSS (NM_029272)	-284	Ndufs7
chr3-11	chr3	51162205	51162427	6.05	0.812	Intergenic	tRNA-Pro-CCA|tRNA|tRNA	-17750	Elf2
chr17-22	chr17	23677831	23678053	6.05	0.795	Intergenic	tRNA-Pro-CCG|tRNA|tRNA	23251	Zfp213
chr11-101	chr11	48631891	48632113	5.81	0.75	Intergenic	Intergenic	-1109	Trim41
chr8-66	chr8	97227776	97227998	5.81	0.604	Intergenic	Intergenic	-7745	Pllp
chr17-24	chr17	23687149	23687371	5.81	0.765	Intergenic	tRNA-Arg-CGG|tRNA|tRNA	13933	Zfp213
chr11-100	chr11	48670092	48670314	5.57	0.574	Intergenic	tRNA-Val-GTY|tRNA|tRNA	14645	Irgm1
chr14-8	chr14	44085546	44085768	5.57	0.767	Intergenic	L1Md_A|LINE|L1	-59313	1700001F09Rik
chr6-17	chr6	1.13E+08	1.13E+08	5.33	0.735	intron (NR_027010, intron 1 of 2)	intron (NR_027010, intron 1 of 2)	1016	Gt(ROSA)26Sor
chr8-72	chr8	1.24E+08	1.24E+08	5.33	0.793	promoter-TSS (NM_133765)	promoter-TSS (NM_133765)	-456	Fbxo31
chr9-36	chr9	1.23E+08	1.23E+08	5.33	0.75	promoter-TSS (NM_030692)	promoter-TSS (NM_030692)	-445	Sacm1l
chr13-53	chr13	23555304	23555526	5.08	0.806	intron (NM_013483, intron 3 of 7)	tRNA-Glu-GAG_|tRNA|tRNA	2355	Btn1a1
chr4-61	chr4	10801121	10801343	5.08	0.714	promoter-TSS (NM_026005)	promoter-TSS (NM_026005)	-413	2610301B20Rik
chr13-44	chr13	22112895	22113117	5.08	0.786	Intergenic	Intergenic	-11396	Prss16

Table A.5. (cont.)

chr11-103	chr11	97658324	97658546	5.08	0.714	Intergenic	tRNA-Cys-TGY|tRNA|tRNA	-2551	Lasp1
chr3-12	chr3	30500062	30500284	5.08	0.694	promoter-TSS (NM_001289621)	promoter-TSS (NM_001289621)	-836	Mynn
chr12-9	chr12	5061942	5062164	5.08	0.767	Intergenic	Intergenic	137894	Atad2b
chr13-43	chr13	22022935	22023157	4.84	0.656	Intergenic	tRNA-Met-i|tRNA|tRNA	-13917	Zfp184
chr13-33	chr13	22009009	22009231	4.84	0.677	Intergenic	Intergenic	-20202	Mir1983
chr13-42	chr13	21418469	21418691	4.84	0.688	Intergenic	Intergenic	14508	Gpx6
chr10-23	chr10	92915677	92915899	4.84	0.657	promoter-TSS (NM_008517)	promoter-TSS (NM_008517)	-353	Lta4h
chr6-19	chr6	1.29E+08	1.29E+08	4.84	0.8	Intergenic	4.5SRNA|scRNA|scRNA	-10829	Klrb1c
chr4-120	chr4	1.56E+08	1.56E+08	4.84	0.846	Intergenic	CpG	-1663	Klhl17
chr13-32	chr13	21947760	21947982	4.84	0.724	Intergenic	Intergenic	22259	Hist1h2bq
chr17-40	chr17	26011959	26012181	4.84	0.821	promoter-TSS (NM_026686)	promoter-TSS (NM_026686)	-375	0610011F06Rik
chr15-4	chr15	36723813	36724035	4.84	0.771	3' UTR (NM_001253805, exon 1 of 6	5' UTR (NM_001253805, exon 1 of 6)	369	Ywhaz
chr13-52	chr13	22120981	22121203	4.84	0.846	Intergenic	tRNA-Ile-ATT|tRNA|tRNA	6420	Hist1h2ah
chr1-222	chr1	1.27E+08	1.27E+08	4.84	0.75	Intergenic	Intergenic	391245	Actr3
chr5-21	chr5	30237824	30238046	4.84	0.75	intron (NR_105781, intron 1 of 2)	L1Md_F|LINE|L1	-101766	Il6
chrX-79	chrX	1.67E+08	1.67E+08	4.84	0.562	Intergenic	Intergenic	-88599	G530011O06Rik
chr9-40	chr9	1.04E+08	1.04E+08	4.84	0.769	Intergenic	L1ME1|LINE|L1	-54498	Cpne4
chr11-104	chr11	95720414	95720636	4.84	0.657	intron (NR_045187, intron 1 of 1)	CpG	394	4833417C18Rik
chr17-39	chr17	25714713	25714935	4.6	0.724	Intergenic	Intergenic	-1295	Lmf1
chr15-5	chr15	9240637	9240859	4.6	0.694	intron (NM_207216, intron 5 of 6)	intron (NM_207216, intron 5 of 6)	-24605	Ugt3a2
chr3-23	chr3	96961941	96962163	4.6	0.828	promoter-TSS (NM_019800)	promoter-TSS (NM_019800)	-648	Acp6
chr5-26	chr5	1.5E+08	1.5E+08	4.6	0.808	intron (NM_010439, intron 1 of 4)	GC_rich|Low_complexity|Low_complexit	1053	Hmgb1
chr17-49	chr17	25410912	25411134	4.6	0.606	promoter-TSS (NR_027452)	promoter-TSS (NR_027452)	-46	Gm17801
chr10-28	chr10	1.28E+08	1.28E+08	4.6	0.679	promoter-TSS (NM_001114096)	promoter-TSS (NM_001114096)	-308	Mir8105
chr9-39	chr9	64585775	64585997	4.6	0.786	promoter-TSS (NM_017382)	promoter-TSS (NM_017382)	-323	Rab11a
chr9-42	chr9	1.2E+08	1.2E+08	4.36	0.792	intron (NM_153287, intron 1 of 4)	intron (NM_153287, intron 1 of 4)	2643	Csrnp1
chr14-37	chr14	44012031	44012253	4.36	0.864	Intergenic	Intergenic	14202	1700001F09Rik
chrX-94	chrX	1.57E+08	1.57E+08	4.36	0.767	Intergenic	tRNA-Val-GTA|tRNA|tRNA	-1881	Rs1
chr17-60	chr17	45308881	45309103	4.36	0.8	Intergenic	B4A|SINE|B4	261664	Cdc5l
chr16-33	chr16	38252646	38252868	4.36	0.742	ntron (NM_001098404, intron 5 of 8	intron (NM_001098404, intron 5 of 8)	42153	Nr1i2
chr1-213	chr1	51219567	51219789	4.12	0.741	intron (NM_019790, intron 7 of 9)	intron (NM_019790, intron 7 of 9)	-126292	Sdpr
chr18-19	chr18	14114196	14114418	3.87	0.778	intron (NM_145492, intron 2 of 7)	Lx5|LINE|L1	16935	Zfp521

Table A.6. Primers and oligonucleotides used in this study

Gene expression primers	DNA sequence (5'>>3')	Amplicon size	Description
ZSCAN5A-F	ACACAAGAGTATTGACGTAACAGGTGA	498	human ZSCAN5A (Tm: 60.6ºC)
ZSCAN5A-R	GCTTGGAATTACACGTAAACCTCTTCTCG		human ZSCAN5A (Tm: 62.2ºC)
ZSCAN5B-F	AAGAGTCCCACAGATCTGGTGAG	388	human ZSCAN5B (Tm: 59.7ºC)
ZSCAN5B-R	TGCTTAGCTGGGAAAAATACTTAAATGAT TTATTG		human ZSCAN5B (Tm: 61.1ºC)
ZSCAN5C-F	TTCCAAACAGTCCCACAGGGG	468	human ZSCAN5C (Tm: 60.6ºC)
ZSCAN5C-R	CATGCAGCGCATAGGCCTG		human ZSCAN5C (Tm: 59.6ºC)
ZSCAN5D-F	CTGTGGTCAATTTTCTTGGCAAGGA	367	human ZSCAN5D (Tm: 60.4ºC)
ZSCAN5D-R	CCCACCACACCTGTAGGACC		human ZSCAN5D (Tm: 60.1ºC)
hYWHAZ-F	ACTTTTGGTACATTGTGGCTTCAA	95	human internal reference control
hYWHAZ-R	CCGCCAGGACAAACCAGTAT		human YWHAZ (Tm: 58ºC)
mZscan5b-F	AGCCGCAGAGGATAAATGGG	211	mouse Zscan5b (Tm: 58.1ºC)
mZscan5b-R	TGGGCTCATCTCAGGTCTCA		mouse Zscan5b (Tm: 58.4ºC)
mYwhaz-F	TTGATCCCCAATGCTTCGC	88	mouse internal reference control
mYwhaz-R	CAGCAACCTCGGCCAAGTAA		human YWHAZ (Tm: 56.4ºC)

non-tRNA DEG primers	DNA sequence (5'>>3')	Amplicon size	
RMRP-F	AGGCTACACACTGAGGACTCT	192	
RMRP-R	CCTGCGTAACTAGAGGGAGC		

tRNA gene primers (Canella *et al.* 2010)	DNA sequence (5'>>3')	Amplicon size	
tRNA_R-Chr9-F	CTCTGTGGCGCAAATGGATAG	54	tR(TCT):chr9
tRNA_R-Chr9-R	TGACCACACTAGGCTCAG		
tRNA_R-Chr17-F	CTCTGTGGCGCAAATGGATAG	54	tR(TCT):chr17
tRNA_R-Chr17-R	GAATTGCTCTATYCGTCACTAG		
tRNA_I-Chr19-F	TCCAGTGGCGCAATCGGT	57	tI(TAT):chr19
tRNA_I-Chr19-R	ATTGCTCCGCTCGCACTGTC		
tRNA_Y-Chr2-F	CCTTCGATAGCTCAGTTGGT	54	tY(GTA):chr2
tRNA_Y-Chr2-R	TTGCCACGCCCTATCCA		

siRNA (Qiagen)	siRNA sequence	Size	siRNA Cat. # (Qiagen)
ZSCAN5A_si4	CGAGAAGAGGUUUACGUGUAA	21	SI00779436
ZSCAN5A_si5	AAGCUAGUCAUCCACAAGAGA	21	SI04221826
ZSCAN5B_si1	ACGUGUGCAAUAAAUCAUUUA	21	SI00503300
ZSCAN5D_si2	CAGGAAGAACCUGAACGAGCA	21	SI02804774

ISH probes	RNA sequence	Size	Description
ZSCAN5B-ISH-Probe	CCAGAGUCAGGUCCACAUCCUCCACCCCAA GAGACAGUGAAUGAUGGUUUUUCUCCAAU GUUACCAGAUGACAUAUACAC	80	ISH probe design to detect human *ZSCAN5B*
mZscan5b-ISH-Probe	GAAAACCCACCUGAGAAAGAGAAAGGUGU CUGAGUAAUUCUGGAUCCAGUUUUUUCCC CUAGAAAGACUGAGAUUGAAGUCUUUUCU CAGAGAAAUGGAGGUUGGGUUG	109	ISH probe design to detect mouse *Zscan5b*

Table A.6. (cont.)

EMSA probes	DNA sequence (5'>>3')	Size
STAP2_M2-F	5’biotin-CGGGTCGGACTCCGCCCCTGCTTCTGA-3’	27
STAP2_M2-R	5’-TCAGAAGCAGGGGCGGAGTCCGACCCG-3’	
STAP2_M1-F	5’biotin-CTGACCACGCCCCCGCGCCCACCCTCTT-3’	28
STAP2_M1-R	5’-AAGAGGGTGGGCGCGGGGGCGTGGTCAG-3’	
STAP2_M2+M1-F	5’biotin-CGGGTCGGACTCCGCCCCTGCTTCTGACCACGCCCCCGCGCCCACCCTCTT-3’	51
STAP2_M2+M1-R	5’-AAGAGGGTGGGCGCGGGGGCGTGGTCAGAAGCAGGGGCGGAGTCCGACCCG-3’	

Figure A.1. DNA sequence and plasmid map of doxycycline-inducible ZSCAN5A overexpression construct, ZSCAN5A-Tet-OE

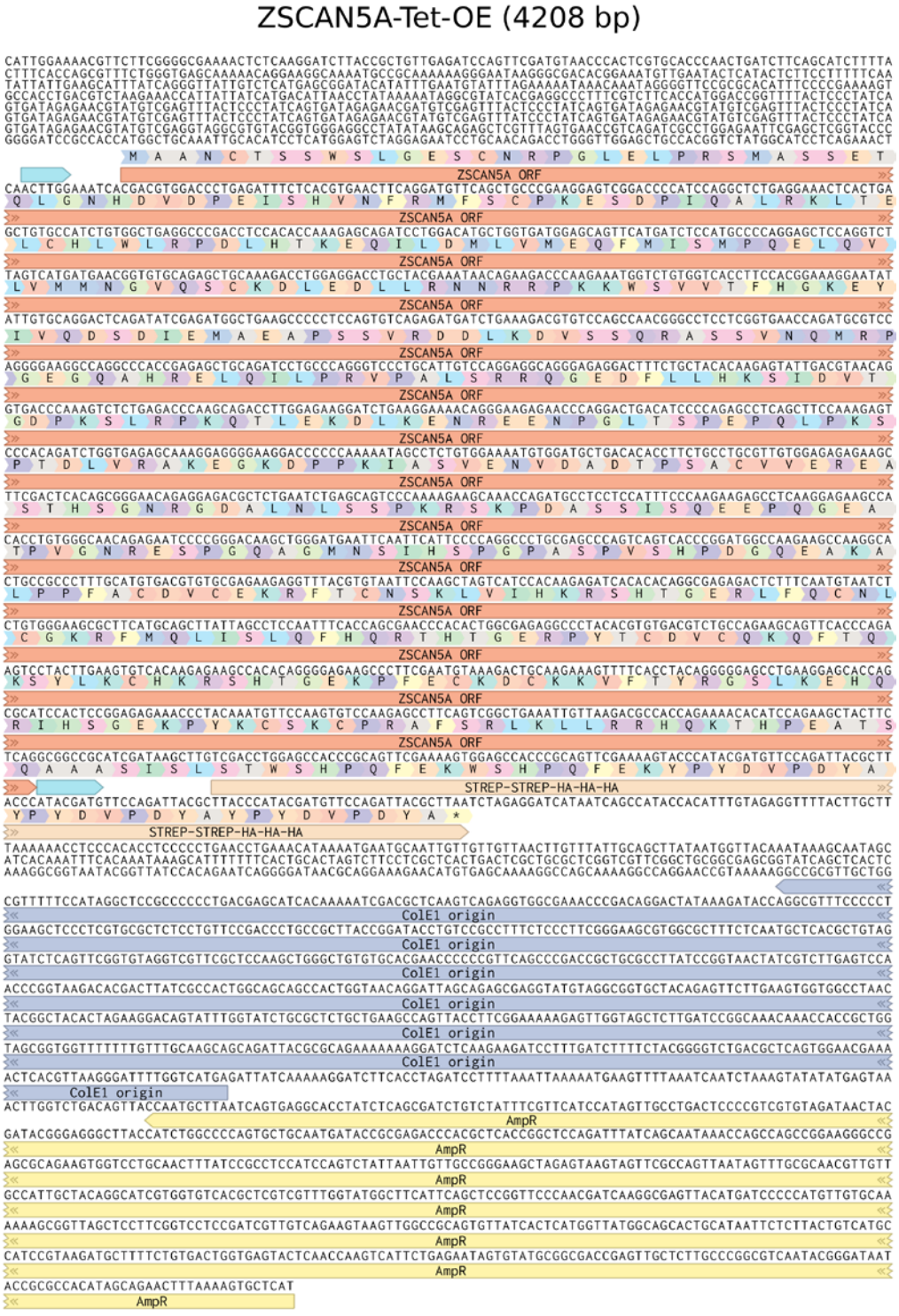


Figure A.1. (cont.)

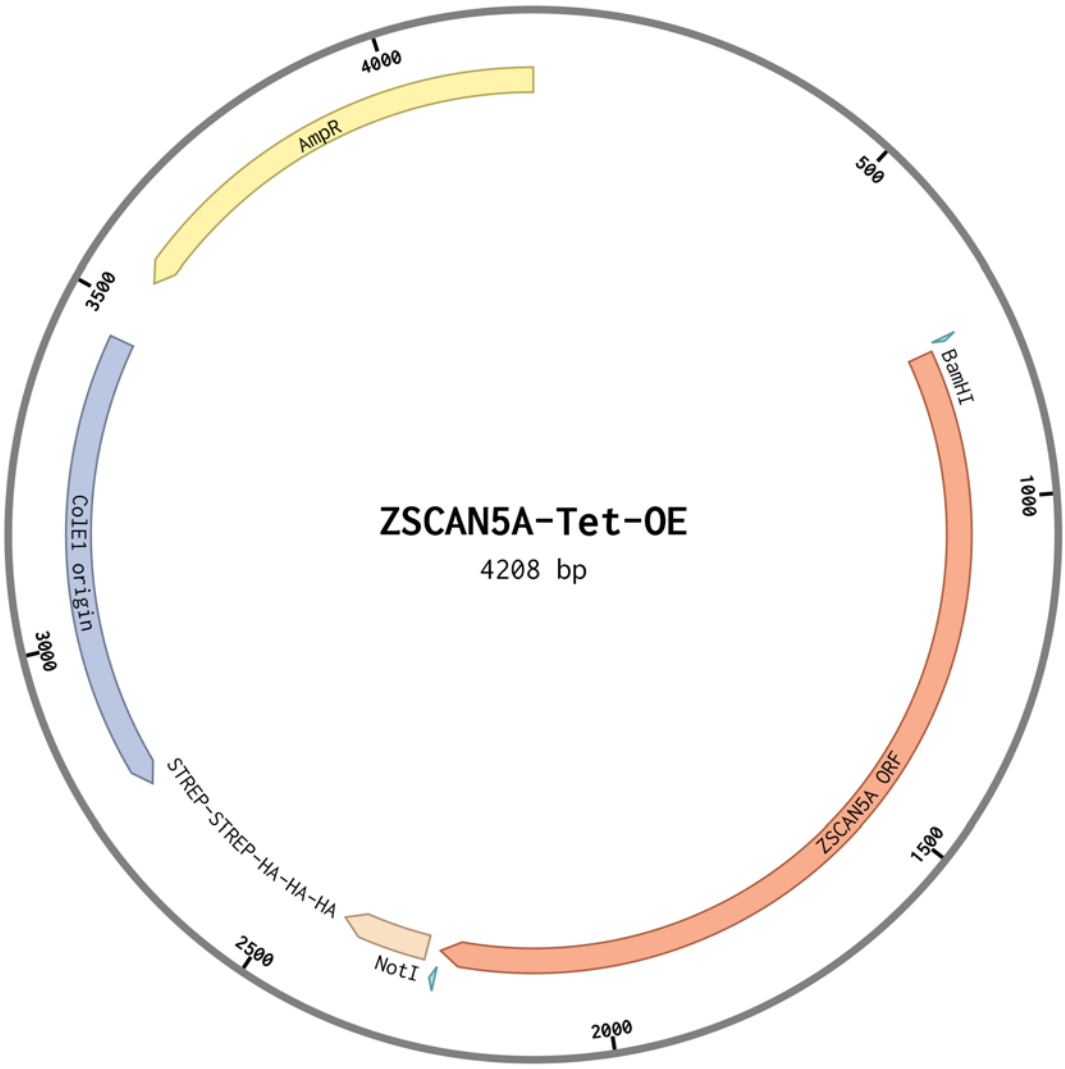


Figure A.2. DNA sequence and plasmid map of doxycycline-inducible ZSCAN5B overexpression construct, ZSCAN5B-Tet-OE

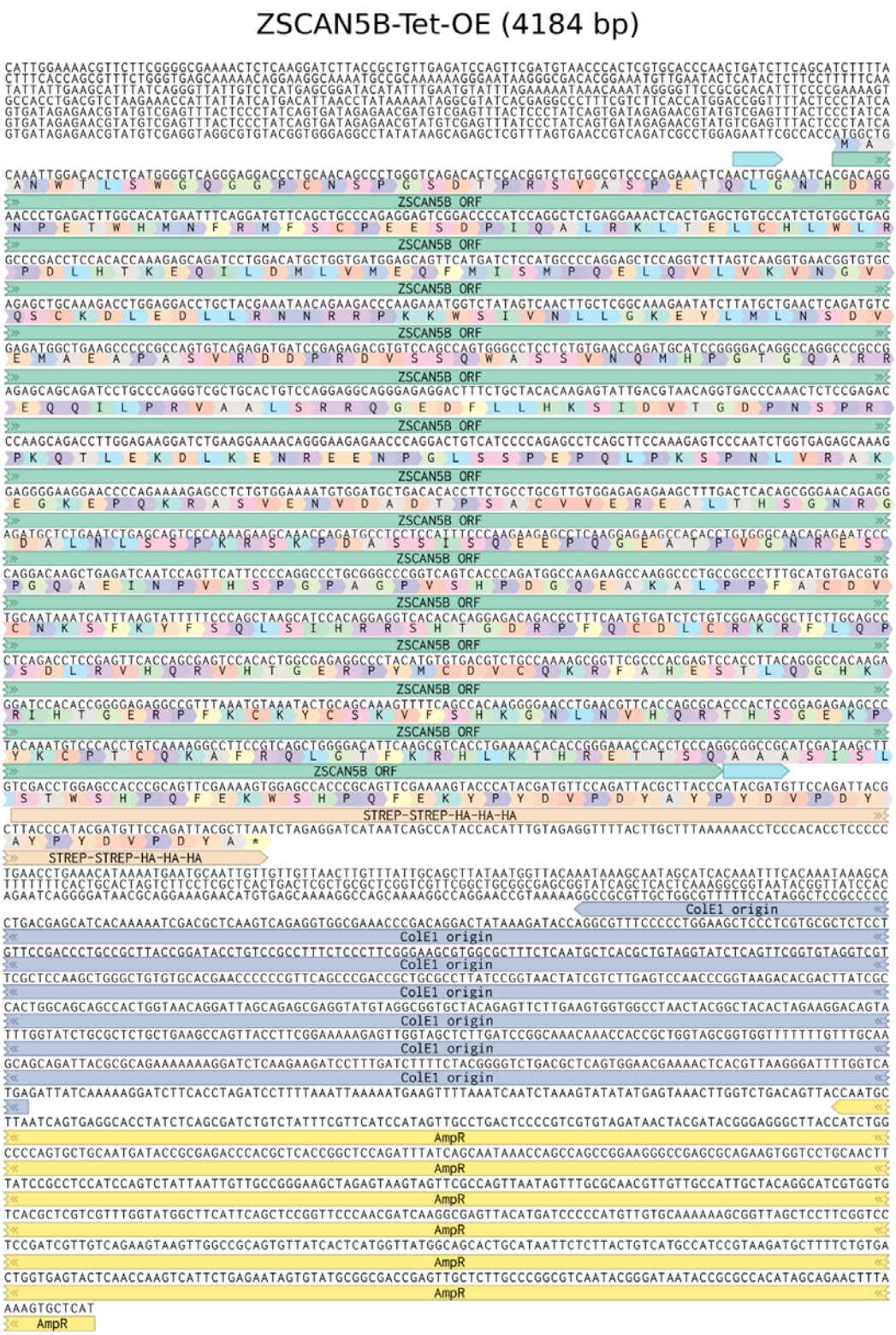


Figure A.2. (cont.)

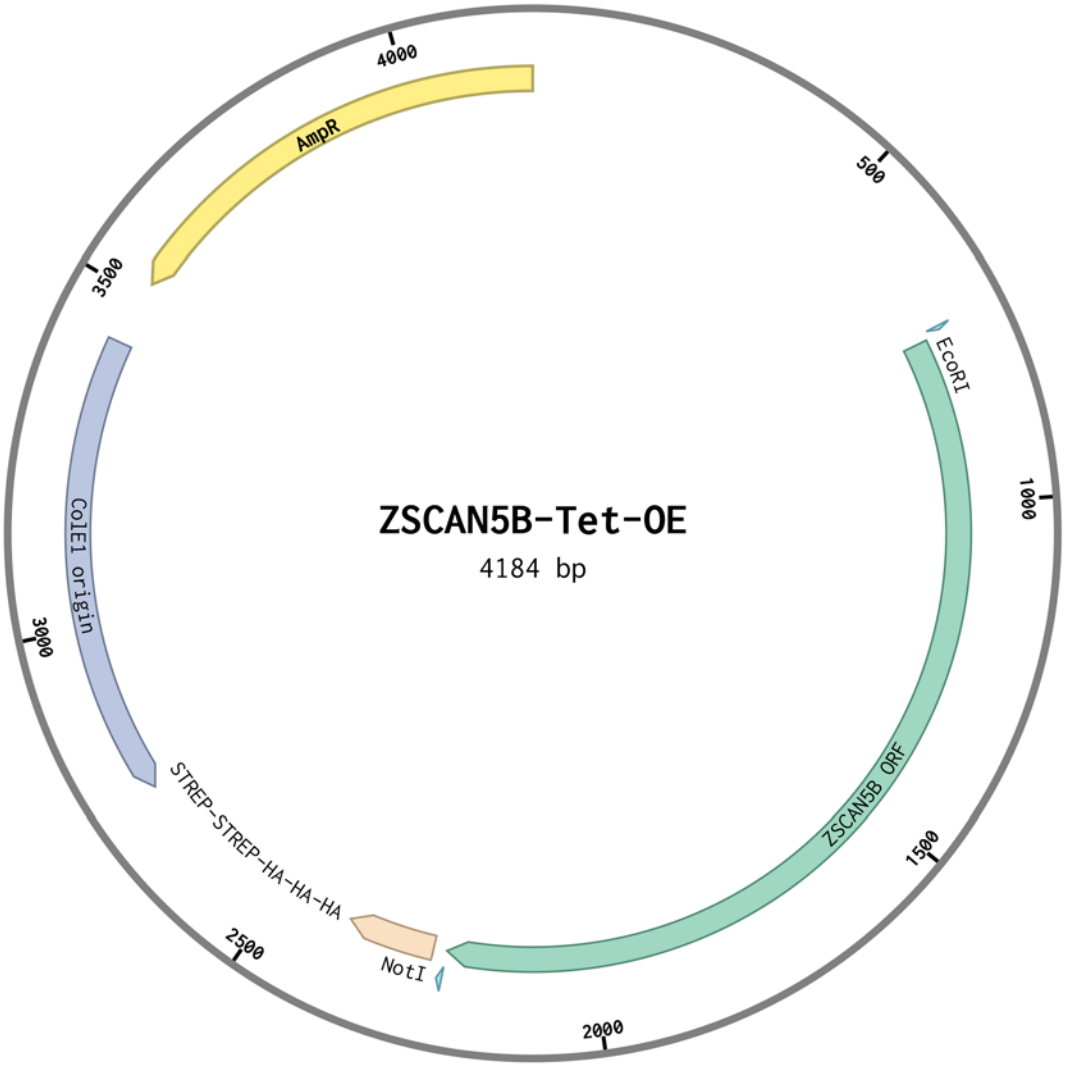


Figure A.3. DNA sequence and plasmid map of doxycycline-inducible ZSCAN5A shRNA construct, ZSCAN5A-Tet-shRNA

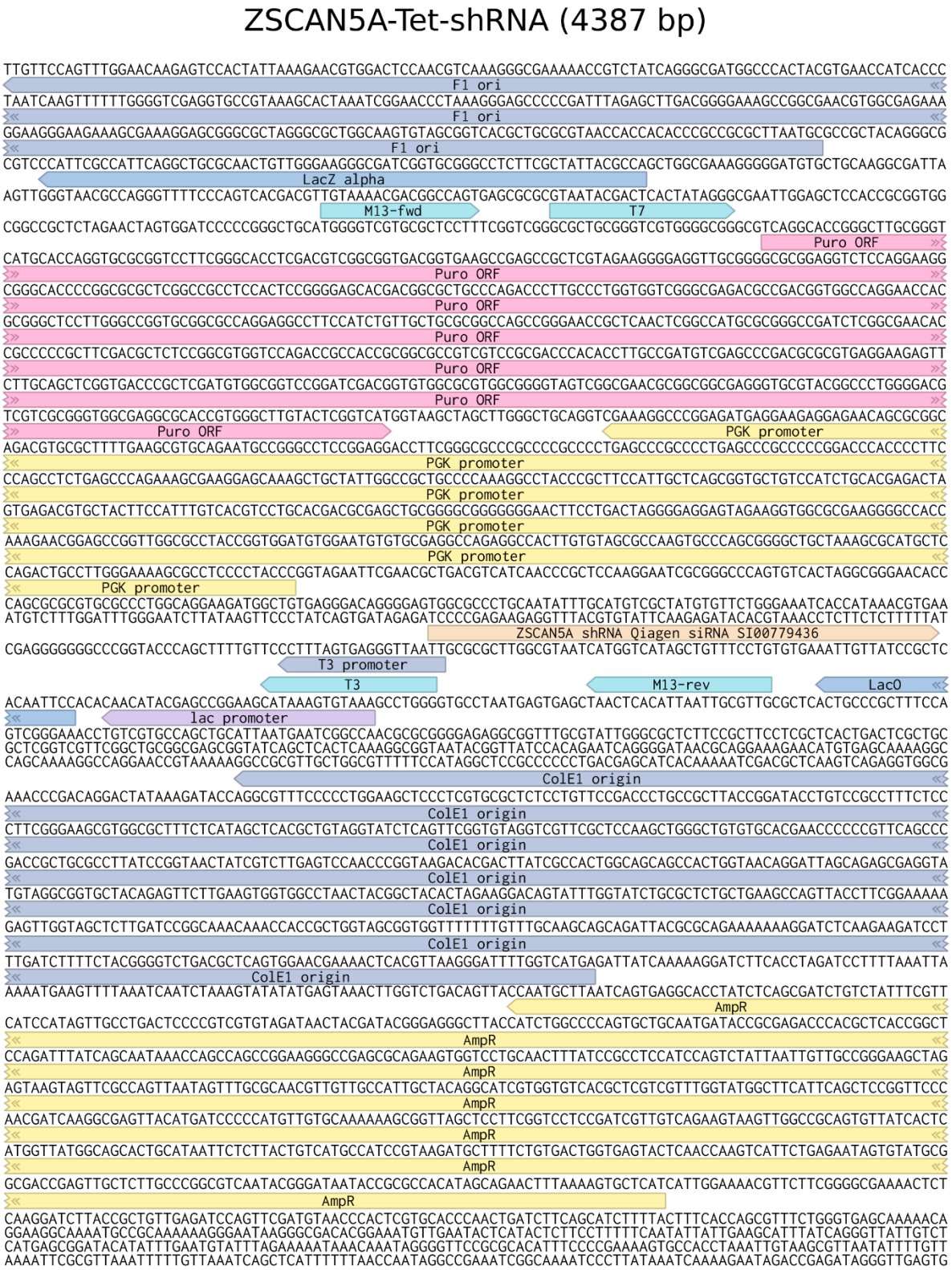


Figure A.3. (cont.)